UUK - 401589-90 - P03891

The American Economic Review

ARTICLES

371

JUNE 1986

THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

Officers

President
ALICE M. RIVLIN
The Brookings Institution

President-Elect
GARY S. BECKER
versity of Chicago

Vice Presidents
PETER A. DIAMOND
Massachusetts Institute of Technology
MANCUR OLSON, JR.
University of Maryland

Secretary
C. ELTON HINSHAW
Vanderbilt University

Treasurer
RENDIGS FELS
Vanderbilt University

Managing Editor of The American Economic Review
ORLEY ASHENFELTER
Princeton University

Managing Editor of The Journal of Economic Literature
JOHN PENCAVEL
Stanford University

Executive Committee

Elected Members of the Executive Committee
VICTOR R. FUCHS
Stanford University
JANET L. NORWOOD
Bureau of Labor Statistics
ALAN S. BLINDER
Princeton University
DANIEL L. McFADDEN
Massachusetts Institute of Technology
SHERWIN ROSEN
University of Chicago
THOMAS J. SARGENT
University of Minnesota

EX OFFICIO Members
CHARLES L. SCHULTZE
The Brookings Institution
CHARLES P. KINDLEBERGER
Massachusetts Institute of Technology

• Printed at Banta Company, Menasha, Wisconsin. The publication number is ISSN 0002-8282.

• *THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the *Survey*, and Supplements, is published by the American Economic Association and is sent to all members five times a year: March; May; June; September; December.

Regular member dues (nonrefundable) for 1986, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature* are as follows:

- $37.50 if annual income is $30,000 or less;
- $45.00 if annual income is above $30,000, but no more than $40,000;
- $52.50 if annual income is above $40,000;
- $18.75 annually for registered students (three years only).

Nonmember subscriptions will be accepted only for both journals: Institutions (libraries, firms, etc.), $105 a year; individuals, $70.00. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add $12.00 to the annual rates above to cover extra postage.

• Correspondence relating to the *Survey*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

• Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

• Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786.

391 .C01

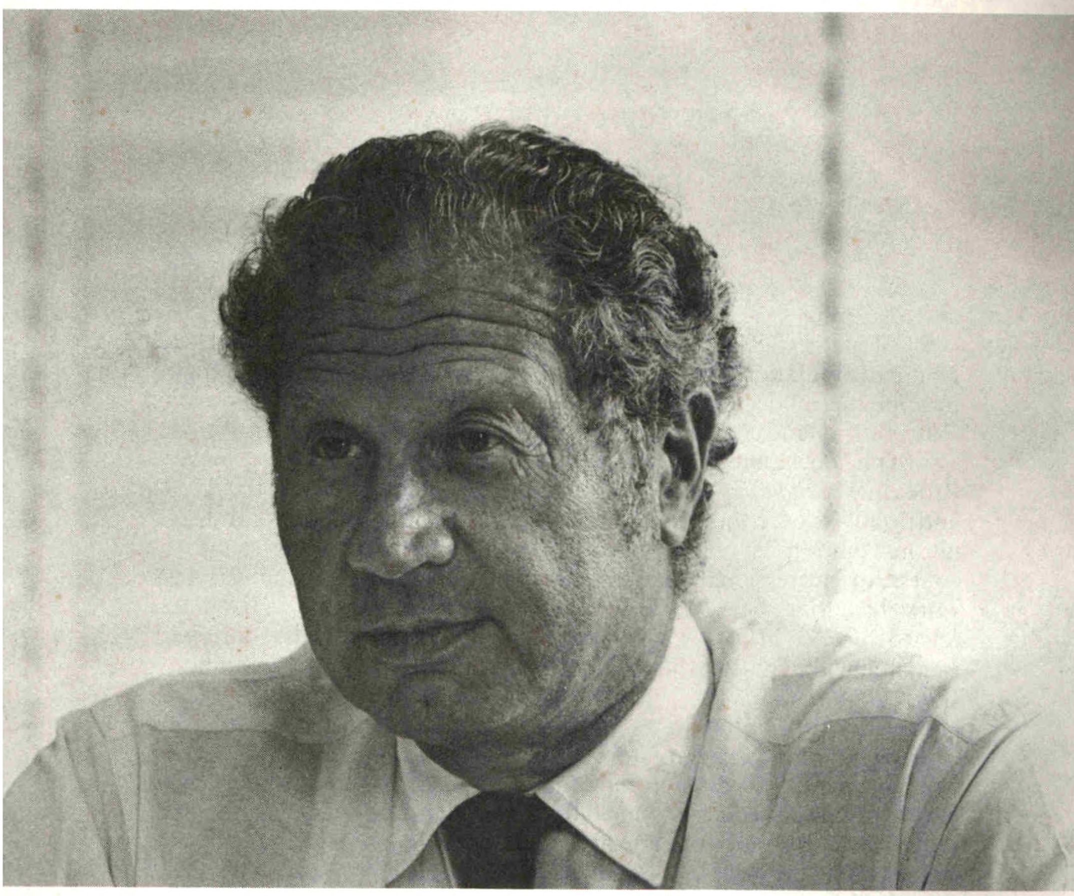
JOSEPH A. PECHMAN

DISTINGUISHED FELLOW

1985

For four decades, as author, editor, research director, and developer of microsimulation files, Joseph A. Pechman has been a leading figure in applying modern economic analysis to the fields of taxation, income distribution, and Social Security policy.

Joseph Pechman helped conceive and then directed the massive Brookings *Studies of Public Finance*, whose output during the 1960's and 1970's—over thirty-five professional volumes, fifty doctoral dissertations, and a large number of journal articles—helped shape the course of professional thought about public finance. Pechman's own work has long been a center of debate on tax policy. His highly regarded *Federal Tax Policy* is in its fourth edition. His study with Benjamin Okner, *Who Bears the Tax Burden*? and its recent updating *Who Paid the Taxes, 1966–1985*?, is the standard analysis of tax burdens in the United States. He has most recently turned his attention to a comparative study of foreign tax systems. Pechman's pioneering work in helping to develop microdata sets from tax and other records has substantially enlarged the ability of scholars and officials to analyze in detail the revenue consequences and distributional effects of changes in tax laws.



Joseph A. Pechman

THE AMERICAN ECONOMIC REVIEW

June 1986

VOLUME 76, NUMBER 3

• Submit manuscripts (4 copies), no longer than 50 pages, double spaced, to:
Orley Ashenfelter, Managing Editor, *AER*; 209 Nassau Street, Princeton, NJ 08542–4607.

• Submission fee: $25 for members; $50 for nonmembers. Style guides will be provided upon request.

Articles

Shorter Papers

Editor's Note

At the meeting of the Executive Committee of the American Economic Association, March 21, 1986,

It was voted to adopt the following policy of disclosure for the Association's journals:
Authors are expected to reveal the sources of any financial or research support received in connection with the preparation of their article.

Life Cycle, Individual Thrift, and the Wealth of Nations

By FRANCO MODIGLIANI*

This paper provides a review of the theory of the determinants of individual and national thrift that has come to be known as the Life Cycle Hypothesis (*LCH*) of saving. Applications to some current policy issues are also discussed.

Section I deals with the state of the art on the eve of the formulation of the *LCH* some thirty years ago. Section II sets forth the theoretical foundations of the model in its original formulation and later amendment, calling attention to various implications, distinctive to it and, sometimes, counterintuitive. It also includes a review of a number of crucial empirical tests, both at the individual and the aggregate level. Section III reviews some applications of *LCH* to current policy issues, though only in sketchy fashion, as space constraints prevent fuller discussion.

I. Antecedents

A. *The Role of Thrift and the Keynesian Revolution*

The study of individual thrift and aggregate saving and wealth has long been central to economics because national saving is the source of the supply of capital, a major factor of production controlling the productivity of labor and and its growth over time. It is because of this relation between saving and productive capital that thrift has traditionally been regarded as a virtuous, socially beneficial act.

Yet, there was a brief but influential interval in the course of which, under the impact of the Great Depression, and of the interpretation of this episode which John Maynard Keynes suggested in the *General Theory* (1936), saving came to be seen with suspicion, as potentially disruptive to the economy and harmful to social welfare. The period in question goes from the mid-1930's to the late 1940's or early 1950's. Thrift posed a potential threat, as it reduced one component of demand, consumption, without systematically and automatically giving rise to an offsetting expansion in investment. It might thus cause "inadequate" demand—and, hence, output and employment lower than the capacity of the economy. This failure was attributable to a variety of reasons including wage rigidity, liquidity preference, fixed capital coefficients in production, and to investment controlled by animal spirits rather than by the cost of capital.

Not only was oversaving seen as having played a major role in the Great Depression, but, in addition, there was widespread fear that the problem might come back to haunt the postwar era. These fears were fostered by a widely held conviction that, in the future, there would not be too much need for additional accumulation of capital while saving would rise even faster than income. This combination could be expected to result, sooner or later, in saving outstripping the "need" for capital. These concerns were at the base of the "stagnationist" school which was prominent in the 1940's and early 1950's.

B. *Early Keynesian Theories of the Determinants of Saving*

It is interesting and somewhat paradoxical that the present day interest and extensive research activity about saving behavior owes its beginnings to the central role assigned by Keynesian economics to the consumption

*Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139.

This article is the lecture Franco Modigliani delivered in Stockholm, Sweden, December 9, 1985, when he received the Nobel Prize in Economic Sciences.

function as a determinant of aggregate demand, and to the concern with oversaving as a source of both cyclical fluctuations and long-run stagnation. It is for this reason that the early endeavor to model individual and aggregate saving behavior was dominated by the views expressed on this subject by Keynes in the *General Theory*, and in particular by his well-known "fundamental psychological [rather that 'economic'] law" (p. 96) to the effect that an increase in income can be counted on to lead to a positive but smaller change in consumption. Even when the analysis followed the more traditional line of demand theory, it relied on a purely static framework in which saving was seen as one of the many "goods" on which the consumer could spend his income. Thus, income was seen as the main systematic determinant of both individual and national saving, and, in line with Keynes' "law," it was regarded as a superior commodity (i.e., one on which "expenditure" rises with income) and most likely a luxury, for which expenditure rises faster than income. Also, in contrast to other goods, the "expenditure" on saving could be negative—and, accordingly, dissaving was seen as typical of people or countries below some "break-even" level of income. All these features could be formalized by expressing consumption as a linear function of income with a substantial positive intercept. This formulation appeared to be supported by the findings of numerous budget studies, and even by the newly developed National Income Accounts, spanning the period of the Great Depression, at the bottom of which saving turned small or even negative.

As is apparent, in this early phase the dominant approach could best be characterized as crudely empirical; little attention was given to why rational consumers would choose to "allocate" their income to saving. The prevailing source of substantial saving was presumably the desire of the rich to bequeath an estate (Keynes' "pride" motive, p. 108). Accordingly, the main source of the existing capital *stock* could be traced to inheritance. Similarly, there was little evidence of concern with how, and how long, "poor" people, or countries, could dissave without having saved first or without exceeding their means.

C. *Three Landmark Empirical Studies*

In the second half of the 1940's, three important empirical contributions dealt a fatal blow to this extraordinarily simple view of the saving process. First, the work of Simon Kuznets (1946) and others provided clear evidence that the saving ratio had not changed much since the middle of the nineteenth century, despite the large rise in per capita income. Second, a path-breaking contribution of Dorothy Brady and R. D. Friedman (1947) provided a reconciliation of Kuznets' results with budget study evidence of a strong association between the saving rate and family income. They demonstrated that the consumption function implied by family data shifted up in time as mean income increased, in such a way that the saving rate was explained not by the *absolute* income of the family but rather by its income *relative* to overall mean income.

Ways of reconciling these findings with the standard linear consumption function were soon provided by James Duesenberry (1949) and me (1949), though within the empirical tradition of the earlier period. Duesenberry's "relative income hypothesis" accounted for the Brady-Friedman results in terms of imitation of the upper classes. This is an appealing explanation, though it fails to come to grips with the budget constraint in the case of would-be dissavers below mean income. Similarly, the "Duesenberry-Modigliani" consumption function tried to reconcile the cyclical variations of the saving ratio with its long-run stability by postulating that current consumption was determined not just by current income but also by its highest previous peak, resulting in a ratchet-like upward creep in the short-run consumption function. In my own formulation, primary stress was placed on reasons why the saving rate should move procyclically and on the consideration that in an economy with stable long-run growth, the ratio of the current to highest previous income could be taken as a good measure of cyclical conditions.

Duesenberry, on the other hand, put more stress on consumers explicitly anchoring their consumption on the previous peak. This formulation was brought to its logical conclusion by Tillman Brown (1952) when he proposed that the highest previous income should be replaced by the highest previous consumption.

The third fundamental contribution was the highly imaginative analysis of Margaret Reid (unpublished) which pointed to a totally different explanation for the association between the saving ratio and relative income, namely that consumption was controlled by normal or "permanent," rather than current, income.

This contribution was an important source of inspiration, both for the life cycle and for the roughly contemporaneous Permanent Income Hypothesis (*PIH*) of Milton Friedman (1957).

II. The Life Cycle Hypothesis

Between 1952 and 1954, Richard Brumberg and I wrote two essays, "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data" (1954), and "Utility Analysis and the Aggregate Consumption Function: An Attempt at Integration" (published in 1979) which provide the basis for the Life Cycle Hypothesis of Saving (*LCH*). They will be referred to hereafter as MB-C and MB-A, respectively. Our purpose was to show that all the well-established empirical regularities could be accounted for in terms of rational, utility-maximizing, consumers allocating optimally their resources to consumption over their life, in the spirit of Irving Fisher (1930). (For an earlier and extensive, but strictly theoretical, application of utility maximization to the theory of saving by households, see U. Ricci, 1926.)

A. *Utility Maximization and the Role of Life Resources* (*Permanent Income*)

The hypothesis of utility maximization (and perfect markets) has, all by itself, one very powerful implication—the resources that a representative consumer allocates to consumption at any age, t, will depend only on his life resources (the present value of labor income plus bequests received, if any) and not at all on income accruing currently. When combined with the self-evident proposition that the representative consumer will choose to consume at a reasonably stable rate, close to his anticipated average life consumption, we can reach one conclusion fundamental for an understanding of individual saving behavior, namely that the size of saving over short periods of time, like a year, will be swayed by the extent to which current income departs from average life resources.

This conclusion is common to *LCH* and to Friedman's *PIH* which differs from *LCH* primarily in that it models rational consumption and saving decisions under the "simplifying" assumption that life is indefinitely long. Accordingly, the notion of life resources is replaced by that of "permanent income," while the discrepancy between current and permanent income is labeled "transitory" income.

The notion that saving largely reflects transitory income has a number of implications which have been made familiar by the contributions of Friedman and by our own 1954 paper, and which have received ample empirical support, even with some occasional controversy. Among these implications, the best known and well established is that relating to the upward bias arising in estimating the slope of a saving-income relation from budget data, when, as is usual, the individual observations are classified by current income classes. Because of the correlation between transitory and current income (relative to mean income), the regression line tends to be steeper than the underlying true relation between the (permanent) saving rate and permanent income. Thus, the estimated saving function departs from the true one by being rotated counterclockwise around the mean, to an extent that is greater the greater the variability of transitory income, for example, more for a sample of farmers than for one of government employees. It is this phenomenon that accounts for the finding of

Brady-Friedman cited above, to the effect that the saving ratio, estimated from budget studies at different points of time, appears to depend on the income not in absolute terms but rather relative to overall mean income.

This same consideration provides an explanation for a famous counterintuitive empirical finding first observed in a large survey conducted in the United States in 1936, namely that black families appeared to save more (or dissave less) than white families at any level of income. The reason, of course, is that black families tend to have a much lower average level of permanent income, and, therefore, at any given level of *current* income the transitory component, and hence saving, tended to be larger (see, for example, Fisher and Brown).

The extent of bias in the cross-sectional saving function should tend to decline if the households are classified by some criterion less positively correlated with transitory income, and this prediction too has been extensively verified (see, for example, my paper with Albert Ando, 1960).

However, I do not intend to pursue here any further the implications of the relation between saving and transitory income since, as already noted, these implications are basically the same for *LCH* as for *PIH*. I concentrate, instead, on those aspects that are specific to *LCH*.

B. *LCH—The "Stripped Down" Version*

By explicitly recognizing the finite life of households, the *LCH* could deal with variations in serving other than those resulting from the transitory deviations of income from life resources of *PIH*. In particular, it could focus on those systematic variations in income and in "needs" which occur over the life cycle, as a result of maturing and retiring, and of changes in family size—hence the name Life Cycle Hypothesis. In addition, the *LCH* was in a position to take into account bequests and the bequest motive, which were not amenable to analysis within the approximation of infinite life.

In MB-C and in the first two parts of the MB-A, we made a number of simplifying, stylized, assumptions concerning the life cycle path of household opportunities and tastes, in order to draw out succinctly the essential implications of the *LCH* approach. These were: 1) opportunities: income constant until retirement, zero thereafter; zero interest rate; and 2) preferences: constant consumption over life; no bequests.

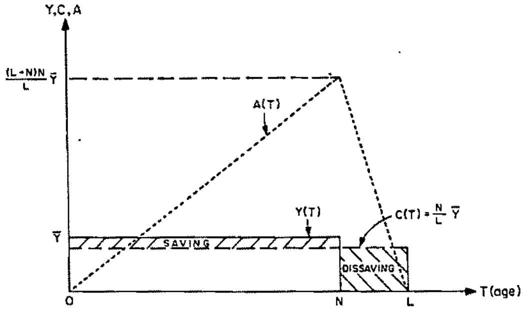


FIGURE 1

For this "basic" or "stripped down" model, the life cycle path of saving and wealth is described in the, by now familiar, graph of Figure 1. Because the retirement span follows the earning span, consumption smoothing leads to a humped-shaped age path of wealth holding, a shape that had been suggested earlier by Roy Harrod (1948) under the label of hump saving (though "hump wealth" would seem like a more descriptive label).

In MB-A, it was shown that this basic model led to a number of implications which were at that time quite novel and surprising—almost counterintuitive. They included the following:

1. The saving rate of a country is entirely independent of its per capita income.

2. Differing national saving rates are consistent with an identical individual life cycle behavior.

3. Between countries with identical individual behavior, the aggregate saving rate will be higher the higher the long-run growth rate of the economy. It will be zero for zero growth.

4. The wealth-income ratio is a decreasing function of the growth rate, thus being largest at zero growth.

5. An economy can accumulate a very substantial stock of wealth relative to income even if no wealth is passed on by bequests.

6. The main parameter that controls the wealth-income ratio and the saving rate for given growth is the prevailing length of retirement.

To establish these propositions, we begin by considering the case of a stationary economy, and then that of steady growth.

1. *The Case of a Stationary Economy.* Suppose that there is neither productivity nor population growth, and assume, conveniently, that mortality rate is 1 at some age L and 0 before. Then, clearly, Figure 1 will represent the age distribution of wealth, saving, consumption, and income, up to a factor representing the (constant) number of people in each age bracket. Hence, the aggregate wealth-income ratio, W/Y, is given by the ratio of the sum of wealth held at each age—the area under the wealth path—to the area under the income path. This has a number of significant implications.

(a) It is apparent from the graph that W/Y depends on a single parameter, the length of retirement, M—which establishes Proposition 6. The relation between M and W/Y turns out to be extremely simple, to wit:

$$W/Y = M/2, \tag{1}$$

(see MB-A, fn. 38).

(b) In MB-A, for illustrative purposes, we conservatively took the average length of retirement as 10 years, implying a wealth-income ratio of 5. This result was an exciting one in that this value was close to the income ratio suggested by preliminary estimates of Raymond Goldsmith's (1956) monumental study of U.S. savings. It implied that one could come close to accounting for the entire wealth holding of the United States without any appeal to the bequest process—Proposition 5—a quite radical departure from conventional wisdom.

(c) With income and population stationary, aggregate wealth must remain constant in time and, therefore, the change in wealth or rate of saving must be zero, despite the large stock of wealth—Proposition 3. The explanation is that, in stationary state, the dissaving of the retired, from wealth accumulated earlier, just offsets the accumulation of the active population in view of retirement. Saving could occur only transiently if a shock pushed W away from $(M/2)\bar{Y}$, where $\bar{Y}$ is the stationary level of income; then, as long as Y remained at $\bar{Y}$, wealth would gradually return to the equilibrium level $(M/2)\bar{Y}$.

2. *The Case of a Steadily Growing Economy.* In this case, the behavior of the saving rates can be inferred from that of aggregate private wealth, W, through the relation $S = \Delta W$, implying:

$$s \equiv \frac{S}{Y} = \frac{\Delta W}{W}\frac{W}{Y} = \rho w \tag{2}$$

$$ds/d\rho = w + \rho(dw/d\rho)$$

where w is the wealth-income ratio and ρ is the rate of growth of the economy which in steady state equals the rate of growth of wealth, $\Delta W/W$. Since w is positive and is based on a level life cycle consumption and earnings, which insures that it is independent of the *level* of income, we have established Propositions 1 and 2. If, in addition, the age profile of the wealth-income ratio could be taken as independent of *growth*, then the saving rate would be proportional to growth with a proportionality factor equal to $M/2$, substantiating Proposition 3. Actually, the model implies that w is, generally, a declining function of ρ—Proposition 4—though with a small slope, so that the slope of the relation between s and ρ tends to flatten out as ρ grows.

When the source of growth is population, the mechanism behind positive saving may be labelled the Neisser effect (see his 1944 article): younger households in their accumulation phase account for a larger share of population, and retired dissavers for a smaller share, than in the stationary society. However, w also falls with ρ because the younger people also are characterized by relatively lower levels of wealth holding. Thanks to the simplifying assumptions of the basic model,

it was possible to calculate explicitly values for w and s: for $\rho = 2$ percent, $w = 4$, $s = 8$ percent; for $\rho = 4$ percent, $w = 3\text{-}1/4$, $s = 13$ percent.

When the growth is due to productivity, the mechanism at work may be called the Bentzel (1959) effect (who independently called attention to it). Productivity growth implies that younger cohorts have larger lifetime resources than older ones, and, therefore, their savings are larger than the dissaving of the poorer, retired cohorts. It was shown in MB-A that, if agents plan their consumption as though they did not anticipate the *future* growth of income, then $w(\rho)$ and $s(\rho)$ for productivity growth are just about the same as for population growth, for values of ρ in the relevant range.

It should be noted that this conclusion is diametrically opposite to that reached by Friedman, namely that productivity growth should tend to *depress* the saving ratio on the ground that a rise in income "expected to continue tends to raise permanent income relative to measured income and so to raise consumption relative to measured income" (p. 234). This difference in the implications of the two models—one of the very few of any significance—can be traced to the fact that, if life is infinite, there cannot be a Bentzel effect. To be sure, to the extent that agents anticipate fully future income, they will tend to shift consumption from the future to the present and this will tend to reduce the path of wealth and perhaps even generate negative net worth in early life (see, for example, James Tobin, 1967). But this effect must be overshadowed by the Bentzel effect, at least for small values of ρ which, realistically, is what matters. (This follows from the continuity of $ds/d\rho$ in equation (2).)

The model also implies that the short-run behavior of aggregate consumption could be described by a very simple aggregate consumption function, linear in aggregate (labor) income (YL), and wealth (W):

$$C = \alpha YL + \delta W. \tag{3}$$

An equation of this type had been proposed somewhat earlier by Gardner Ackley (1951), though both the functional form and the presumed stability of the coefficients rested on purely heuristic considerations. By contrast, it was shown in MB-A that, if income followed closely the steady growth path, then the parameters α and δ could be taken as constant in time and determined by the length of life (L), of retirement (M), and the rate of growth (MB-A, p. 135). For the standard assumption $L = 50$, $M = 10$ and $\rho = .03$, δ comes to .07 (see MB-A, p. 180). Furthermore, the parameters could be well approximated by the same constant even if income moved around the trend line, as long as the departures were not very long lasting and deep, except that YL should be interpreted as long-run expected rather than current income. The short-run equation (3) is, of course, consistent with the long-run properties 1 to 6, as one can readily verify.

3. *Empirical Verifications.* None of these long- and short-run implications of the basic model could be explicitly tested at the time they were established. There were no data on Private Net Worth to test equation (3), except for some indirect estimates pieced together by W. Hamburger (1951) and some preliminary Goldsmith figures for a few selected years. Similarly, information on Private National Saving were available only for a couple of countries. We could only take encouragement from the fact that the model seemed to fit the single observation available, namely the United States. Both the wealth-income ratio, 4 to 5, and the saving rate, S, "between 1/7 and 1/8" (Goldsmith) were broadly consistent with the prediction of the model, for a 3 percent growth rate, namely 4-1/3 for w and 13 percent for s.

But the availability of data improved dramatically in the next decade. For the United States an annual time-series of Private Wealth was put together in the early 1960's (Ando et al., 1963), and equation (3) was tested (my article with Ando, 1963). It was found to fit the data quite well, and with parameter estimates close to those predicted by the model. By now the consumption function (3) has become pretty much standard,

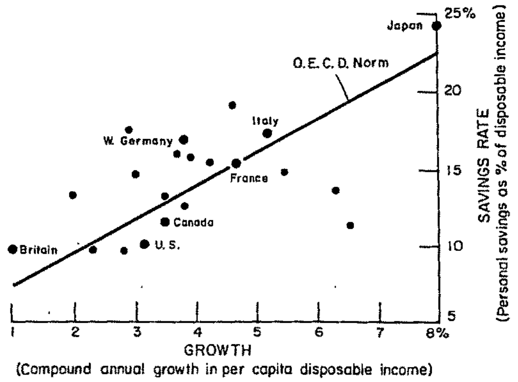


FIGURE 2

Source: My paper with A. Sterling (1983).

having been estimated for many countries and periods. The coefficient of wealth is frequently lower than .07 quoted earlier but this can be accounted for, at least in part, by the fact that Y is typically defined as total rather than just labor income.

Similarly, by the early 1960's, the United Nations had put together National Account statistics for a substantial number of countries, characterized by wide differences in the growth rate, and it became possible to test the relation between the national saving ratio and the growth rate. The early tests were again quite successful (Hendrik Houthakker, 1961 and 1965; Nathaniel Leff, 1969; and myself, 1970). The newly available data also revealed the puzzling and shocking fact that the saving ratio for the United States, by far the richest country in the world, was rather low compared with other industrial countries (see, for example, Figure 2). The *LCH* could account for the puzzle through a relatively modest growth rate. By now it is generally accepted that growth is a major source of cross-country differences in the saving rate.

C. *The Effect of Dropping the Simplifying Assumptions*

As was demonstrated in MB-A, most of the simplifying assumptions can be replaced by more "realistic" ones without changing the basic nature of the results, and, in particular, the validity of Propositions 1 to 5.

1. *Nonzero Interest.* Allowing for a nonzero interest rate, r, has two effects. One effect is on income as we must distinguish between labor income, say YL, property income, YP, whose "permanent component" may be approximated by rW, and total income, $Y = YL + YP = YL + rW$. If we continue to assume a constant labor income till retirement, then the graph of income in Figure 1 is unchanged. However, the graph of consumption changes through an income and substitution effect: the addition of rW increases income, but at the same time r also affects the opportunity cost of current, in terms of future consumption. It is possible that the consumer would still choose a constant rate of consumption over life (if the elasticity of substitution were zero). In this case, in Figure 1, consumption will still be a horizontal straight line, but at a higher level because of the favorable "income effect" from rW. As for saving, it will be the difference between C and Y. The latter differs from the (piecewise) horizontal YL in the figure by rW, which is proportional to W. As a result, the path of W will depart somewhat from the "triangle" of Figure 1, and, in particular, the overall area under the path can be shown to decline with r. This means that W and, a fortiori, $w = W/Y$, will fall with r.

This result has interesting implications for the much debated issue of the effect of interest rates on saving. Turning back to equation (2), we see that: (i) in the absence of growth, a change in r has *no* effect on saving (which remains zero), and (ii) for any positive rate of growth, a higher interest rate means a lower saving rate. However, this conclusion depends on the special assumption of zero substitution. With positive substitution, consumption will start lower and will rise exponentially: this "postponement" of consumption, in turn, lifts saving and peak assets. If the substitution effect is strong enough, w will rise and so will s, as long as ρ is positive.

This same conclusion can be derived from (3) and the definition of Y. These can be

shown to imply

$$W/Y = (1-\alpha)/(\rho+\delta-\alpha r). \tag{4}$$

Numerical calculations in MB-A suggest that α is not much affected by r, but δ is. In my 1975 paper, I hypothesized that the effect of r on δ might be expressed as $\delta = \delta^* + \mu r$ when μ is unity for 0 substitution, and declines with substitution (possibly to a negative value). Substituting for δ in (4), one can see that, when the interest rate rises, saving may fall or rise depending on whether μ is larger or smaller than α.

Which of these inequalities actually holds is an empirical matter. Unfortunately, despite a hot debate, no convincing general evidence either way has been produced, *which leads me to the provisional view that s is largely independent of the interest rate*. It should be noted in this connection that, insofar as saving is done through pension schemes aimed at providing a retirement income, the effect of r on s is likely to be zero (or even positive) in the short run but negative in the long run.

2. *Allowing for the Life Cycle of Earning and Family Size*. Far from being constant, average labor income typically exhibits a marked hump pattern which peaks somewhat past age 50, falls thereafter, partly because of the incidence of retirement, and does not go to zero at any age, though it falls sharply after 65. However, consumption also varies with age, largely reflecting variations in family size, as one might expect if the consumer smooths consumption *per equivalent adult* (my paper with Ando, 1957). Now the life cycle of family size, at least in the United States, has a very humped shape rather similar to that of income, though with a somewhat earlier peak. As a result, one might expect, and generally finds, a fairly constant rate of saving in the central age group, but lower saving or even dissaving in the very young or old. Thus, as in Figure 1, the wealth of a given cohort tends to rise to a peak around age 60 to 65 (see, for example, Dorothy Projector, 1968; M. A. King and L.-D. L. Dicks-Mireaux, 1982; R. B. Avery et al., 1984; Ando and A. Kennickell, 1985; and Peter Diamond and Jerry Hausman, 1984).

It is also worth noting that available evidence supports the *LCH* prediction that the amount of net worth accumulated up to any given age in relation to life resources is a decreasing function of the number of children, and that saving tends to fall with the number of children present in the household and to rise with the number of children no longer present (cf. Alan Blinder, Robert Gordon, and David Wise, 1983; and Ando and Kennickell).

3. *Length of Working and Retired Life*. One can readily drop the assumption that the length of retired life is a given constant. As is apparent from Figure 1, a longer retirement shifts forward, and raises, the peak of wealth, increasing w and the saving rate. This does not affect the validity of Propositions 2 to 6, but could invalidate 1. It is possible, in fact, that, in an economy endowed with greater productivity (and, hence, greater per capita income), households might take advantage of this by choosing to work for fewer years. This, in turn, would result in a higher national saving rate. Note, however, that this scenario need not follow. The increase in productivity raises the opportunity cost of an extra year of retirement in terms of consumables, providing an incentive to *shorter retirement*. Thus the saving rate could, in principle, be affected by per capita income, but through an unconventional life cycle mechanism, and, furthermore, in a direction unpredictable a priori. Empirical evidence suggests that the income effect tends to predominate but is not strong enough to produce a measurable effect on the saving rate (my paper with A. Sterling, 1983).

Aside from income, any other variable that affects the length of retirement could, through this channel, affect saving. One such variable, that has received attention lately, is Social Security. Several studies have found that the availability of Social Security, and terms thereof, can encourage earlier retirement (Martin Feldstein, 1974, 1977; Alicia Munnell, 1974; Michael Boskin and Michael Hurd, 1978; myself and Sterling, 1983; and Diamond and Hausman). To this extent, So-

cial Security tends to encourage saving, though this effect may be offset, and even more than fully, by the fact that it also reduces the need for private accumulation to finance a given retirement.

4. *Liquidity Constraint*. Imperfections in the credit markets as well as the uncertainty of future income prospects may, to some extent, prevent households from borrowing as much as would be required to carry out the unconstrained optimum consumption plan. Such a constraint will have the general effect of postponing consumption and increase w as well as s. But, clearly, these are not essential modifications, at least with respect to the aggregate implications—on the contrary, they contribute to insure that productivity growth will increase the saving rate. However, significant liquidity constraints could affect quantitatively certain specific conclusions, for example, with respect to temporary tax changes (see Section III, Part 1, below).

5. *Myopia*. The *LCH* presupposes a substantial degree of rationality and self-control to make preparations for retired consumption needs. It has been suggested—most recently by H. M. Shefrin and Richard Thaler (1985)—that households, even if concerned in principle with consumption smoothing, may be too myopic to make adequate reserves. To the extent that this criticism is valid, it should affect the wealth-income ratio in the direction opposite to the liquidity constraint, though the effect of transitory changes in income from any source would go in the same direction. However, such myopia is not supported empirically. The assets held at the peak of the life cycle are found to represent a substantial multiple of average income (in the order of 5, at least for the United States) and an even larger multiple of permanent income which, in a growing economy, is less than current income. Such a multiple appears broadly consistent with the maintenance of consumption after retirement. This inference is confirmed by recent studies which have found very little evidence of myopic saving behavior. In particular, both Laurence Kotlikoff et al. (1982) and Blinder et al. (especially Figure 4.1), working with data on households close to retirement, find that for most families the resources available to provide for retired consumption appear to be quite adequate to support retired consumption at a rate consistent with life resources.

D. *The Role of Bequests and the Bequest Motive*

Obviously bequests exist in market economies (and not only in market economies). How does their presence affect the relevance and usefulness of the model, and, in particular, the validity of Propositions 1 to 5? In attacking this problem, one must distinguish the issue of principle from the empirical one of how important a role bequests may play in the accumulation of wealth.

1. *How Important are Bequests in the Accumulation of Wealth*? This is an interesting question. The traditional approach took it for granted that bequests are a major source of the existing wealth, while the *LCH* suggested that they might not contribute appreciably.

I recently (1985) reviewed a substantial body of information on inherited wealth based on direct surveys of households and on various sources of estimates on the flow of bequests. This review yields a fairly consistent picture suggesting that the proportion of existing wealth that has been inherited is around 20 percent, with a margin of something like 5 percentage points.

This conclusion is at odds with that presented in a provocative paper of Kotlikoff and Lawrence Summers (1981, hereafter K-S). They endeavor to estimate the share of bequests by two alternative methods: 1) from an estimated flow of bequests, as above, and 2) by subtracting from an independent estimate of private wealth in a given year, their own estimate of the amount of life cycle wealth, accumulated by every cohort present in that year. Using the first method, K-S reach an estimate of inherited wealth of over one-half, while using the second—which they regard as more reliable—their estimate rises even higher, to above four-fifths. In the 1985 paper, I have shown that the difference between my estimate and their much higher

ones can be traced (*i*) to some explicit errors of theirs, for example, their treatment of the purchase of durable goods, and (*ii*) to unconventional definitions, both of inherited wealth of life cycle saving. I have shown that when one corrects the error and uses the accepted definitions, one of the K-S measures—that based on bequest flows—coincides very closely with all other estimates. Their alternative measure remains somewhat higher, but I show it is subject to an appreciable upward bias which could easily account for the difference.

Kotlikoff and Summers have suggested an alternative operational criterion of "importance" which should be independent of definitional differences, namely: by what percentage would aggregate wealth decline if the flow of bequests declined by 1 percent? The suggestion is sound but is very hard to implement from available observations. Nonetheless, it would appear this effect, measured in terms of its impact through inherited wealth, can be taken as approximately equal to the observed share of bequeathed wealth, when wealth is measured according to the conventional definition. Thus, with either measure, bequeathed wealth can be put at less than 25 percent.

The only other country for which the relevant information is available seems to be the United Kingdom (see Royal Commission, 1977). The estimated share of inherited wealth is, again, close to 20 percent.

2. *The Behavior of Saving and the Wealth of the Aged.* A quite different ground for questioning whether the accumulation of wealth can be better accounted for by a life cycle parable than by a bequest motive is to be found in the behavior of saving and assets of elderly households, especially after retirement. The basic *LCH* implies that, with retirement, saving should become negative, and thus assets decline at a fairly constant rate, reaching zero at death. The empirical evidence seems to reveal a very different picture: dissaving in old age appears to be at best modest (for example, see J. Fisher, 1950; Harold Lydall, 1955; T. W. Mirer, 1979, and Ando and Kennickell). According to Mirer, the wealth-income ratio actually continues to rise in retirement. (Note, however, that his estimate is biased as a result of including education in his regression. Given the steady historical rise in educational levels, there will be a strong association between age, educational attainment, and socioeconomic status *relative* to one's cohort if one holds constant the absolute level of education. Thus, his results could merely reflect the association between bequests, wealth, and relative income discussed below.) Most other recent analysts have found that the wealth of a given cohort tends to decline after reaching its peak in the 60–65 age range (A. F. Shorrocks 1975; King and Dicks-Mireaux; Diamond and Hausman; Avery et al.; Ando, 1985; Hurd, 1986), though there are exceptions—for example, Paul Menchik and Martin David (1983) discussed below. To be sure, the results depend on the concept of saving and wealth used. If one makes proper allowance for participation in pension funds, then the dissaving (or the decline in wealth) of the old tends to be more apparent, and it becomes quite pronounced if one includes an estimate of Social Security benefits. But, when the saving and wealth measures include only cash saving and marketable wealth, the dissaving and the decline appears weaker or even absent. Also, those studies which provide median as well as mean values (for example, Ando, 1985), suggest that the picture of a steady decline in wealth is clearer for the median than for the mean which has a more erratic behavior, reflecting the extreme variability of the data.

There are several considerations that can account, at least partly, for the above finding within an *LCH* framework. In particular, the survey data may give an upward biased picture of the true behavior of wealth during old age for two reasons. First, as Shorrocks has argued, one serious bias arises from the well-known positive association between longevity and (relative) income. This means that the average wealth of successively older age classes is the wealth of households with higher and higher life resources, hence the age profile of wealth is upward biased. Second, in a similar vein, Ando and Kennickell

have found evidence that aged households which are poor tend to double up with younger households and disappear from the sampled population so that the wealth of those remaining independent is again an upward biased estimate of average wealth.

3. *Bequests and Uncertainty of the Length of Life*. While it is difficult to assess the extent of these biases, the decumulation, at least of the marketable assets, would seem to be too slow to be explained by the basic *LCH*. A possible partial reconciliation is provided by giving explicit recognition to the existence of uncertainty about the length of life. Indeed, in view of the practical impossibility of having negative net worth, people tend to die with some wealth, unless they can manage to put all their retirement reserves into life annuities. However, it is a well-known fact that annuity contracts, other than in the form of group insurance through pension systems, are extremely rare. Why this should be so is a subject of considerable current interest. It is still ill-understood. "Adverse selection," causing an unfavorable payout, and the fact that some utility may be derived from bequests (Andre Masson, 1986)—see below—are, presumably, an important part of the answer.

In the absence of annuities, the wealth left behind will reflect risk aversion and the cost of running out of wealth. This point has been elaborated in particular by J. B. Davies (1981) who has shown that, for plausible parameters of the utility function including a low intertemporal elasticity of substitution, the extent to which uncertainty of life depresses the propensity to consume increases with age. As a result, "uncertain life time could provide the major element in a complete explanation of the slow decumulation of the retired" (relative to what would be implied by a standard *LCH* model). This conclusion is reinforced by allowing for the uncertainty of major medical expenses. Note also that the wealth bequeathed as a result of a precautionary motive, related to uncertainty of death, must tend, on the average, to be proportional to life resources. Hence, it can be readily incorporated into the basic model and the result labelled *LCH* cum precautionary bequests.

These considerations may go part way toward explaining the slow decumulation. Still, this phenomenon may also reflect, in part, the working of an explicit bequest motive and life planning for it. We may, therefore, ask whether there is any intrinsic inconsistency between a significant amount of bequests induced by a bequest motive and the *LCH* view of the world, in particular, implications 1 to 5.

4. *Bequest Motive in the LCH*. First, it is obvious that no inconsistency arises if planned bequests are, on average, proportional to life resources. However, this possibility is uninteresting. The most casual observation suggests that the planning and leaving of bequests is concentrated in the upper strata of the distribution of life resources, by which we now mean the sum of (discounted) lifetime labor income and bequests received. This observation suggests the following hypothesis, first proposed in MB-A (pp. 173–74):

HYPOTHESIS I: *The share of its resources that a household earmarks, on the average, for bequests is a (nondecreasing) stable function of the size of its life resources relative to the average level of resources of its age cohort.*

We might expect the share to be close to zero until we reach the top percentiles of the distribution of resources, and then to rise rapidly with income.

One can readily demonstrate (cf. my 1975 article) that this assumption assures that Propositions 1 to 5 will continue to hold, at least as long as:

HYPOTHESIS II: *The frequency distribution of the ratio of life resources to mean life resources for each age group is also stable in time.*

Indeed, under these conditions, if income is constant, wealth will also tend to be constant and, therefore, saving to be zero, even in the presence of bequests. To see this, note

first that Hypothesis I insures that bequests left (BL) are a fraction, say γ, of life resources, $\hat{\gamma}$, $BL = \gamma(\hat{Y} + BR)$, where BR is bequests received. Hypothesis II in turn insures that γ is constant in time (and presumably less than one). Next, note that life savings, LS, is given by

$$(5) \quad LS = BL - BR = \gamma\hat{Y} - (1-\gamma)BR.$$

Thus, LS increases with Y and decreases with BR, and is zero if $BR = [\gamma/(1-\gamma)]\hat{Y}$. But this last condition must hold in long-run equilibrium since, if BR is smaller, then there will be positive saving which will increase BR, and reduce LS toward zero; and *vice versa* if BR is larger.

This generalization of the basic model has a number of implications, a few of which may be noted here.

(i) The age patterns of Figure 1 for a stationary society are modified, as bequests raise the average wealth path by a constant, equal to BR, beginning at the age at which bequests are received. The new path remains parallel to the old so that at death it has height $BL = BR$.

(ii) If labor income is growing at some constant rate, then average BR will tend to grow at this same rate and so will BL, but BL will exceed BR by a factor $e^{\rho T}$, where T is the average age gap between donor and recipient. Thus, with positive growth, and then only, the existence of bequests involves life saving, on top of hump saving. In other words, bequests result in a higher wealth-income ratio, depending on γ, and a higher saving ratio, to an extent that is proportional to ρ.

(iii) The share of life resources left as bequests could be an increasing function of the household's resources *relative* to the resources of his cohort. This, in turn, implies that at any age, the saving-income and wealth-income ratio for individual families could be an increasing function of *relative* (not absolute) income.

This last proposition, which is clearly inconsistent with PIH, is supported by a good deal of empirical evidence, beginning with Brady and Friedman. As for the first part of (iii), and the underlying Hypothesis I, it receives strong support from a recent test by Menchik and David. In this imaginative contribution, the authors have assembled, from probate records, a large body of data on individual bequests which they have matched with income data from tax returns. Their sample covers persons born since 1880 (including a few before) and deceased between 1947 and 1978. They find striking evidence that (a) bequests depend on the position of the household's life resources in the distribution of life resources of *its cohort*, (b) that they are small for people whose estimated life resources fall below the 80th percentile in that distribution but that, (c) beyond the 80th percentile, they rise rapidly with (permanent) income.

5. *The Individual Bequests and the Share of Bequeathed Wealth — A Reconciliation.* There remains one serious puzzle. If something like two-thirds of peak wealth is passed on at death, be this "unintentional" transmission through precautionary saving or the conscious result of a desire to bequeath, how can the share of wealth received by bequests amount to less than 25 percent of the total?

Recent contributions of Kennickell (1984) and Ando and Kennickell have pointed the way to a satisfactory resolution, by demonstrating that, in the presence of significant growth, the share of wealth inherited is *not* a satisfactory indication of the importance of bequests. To understand their argument, suppose, conveniently, that *all* wealth ever accumulated is passed on at death, there being therefore *no* life cycle (hump) saving. If the economy is stationary, and thus saving is zero, it will be true that all wealth is due to the bequest motive. It will also be true that all existing wealth is inherited so that, in this case, the share of bequeathed wealth will provide a valid measure of the importance of bequests. But suppose there is growth. Then there is also saving and, therefore, a portion of the existing wealth will be held by those who are accumulating it on its way to be bequeathed. And that portion rises rapidly with growth: for example, at 3 percent growth, bequests left are, on the average, some 2-1/2 times larger than those received, and, correspondingly, the share of wealth

received by bequests falls to just below 40 percent (Kennickell), even though all wealth would again disappear in the absence of the bequest motive.

The empirical relevance of this conclusion has been confirmed by an interesting calculation carried out by Ando and Kennickell (A-K). Starting from estimates of national saving and allocating them by age, using the saving-age relation derived from a well-known budget study (the Bureau of Labor Statistics' *Consumer Expenditure Survey*, 1972–73), they are able to estimate the aggregate amount of wealth accumulated through life saving by each cohort living in a given year. They then compare this with aggregate wealth to obtain an estimate of the shares of wealth that are, respectively, self-accumulated and inherited.

Even though the age pattern of saving they use involves relatively little dissaving in old age, their estimate of the share of inherited wealth turns out to be rather small. For the years after 1974, it is around 25 percent, which agrees well with, and thus supports, the findings of my 1985 paper. For the years 1960 to 1973, the share they compute is somewhat larger, fluctuating between 30 and 40 percent. But this higher figure may at least partly reflect an upward bias in the A-K estimate of inherited wealth. The bias arises from the fact that the change in overall real wealth includes capital gains, while the change in the self-accumulated portion largely excludes them. In the period before 1974, capital gains were unquestionably significantly positive, and hence self-accumulation is underestimated and the share of bequests overestimated. In the years from 1973 to 1980, depressed conditions in the stock market reduce the significance of this effect, though this is partially offset by the boom in real estate values.

E. *A Summing Up*

We have found that the basic version of the *LCH* has proved quite helpful in understanding and predicting many aspects of individual and aggregate saving and wealth-holding behavior. However, two of the assumptions embodied in the stripped down version—a deterministic length of life and the absence of a bequest motive appear, in the light of presently available information, to be conspicously counterfactual. There is substantial evidence that wealth declines slowly in old age—even after correcting for various sources of bias—implying that households, on the average, leave substantial bequests relative to peak wealth.

This evidence can be readily accommodated within the generalized *LCH* framework. That portion of bequests that arises from the precautionary motive can be handled by a straightforward relaxation of the assumptions to allow for a stochastic length of life and risk-averse behavior. The holding of wealth arising from this mechanism can be rightfully regarded as life cycle wealth since it reflects the optimum allocation of resources to consumption over life. Furthermore, the expected size of bequests relative to life resources should be largely independent of resources. The remaining bequests arising from a genuine bequest motive can also be accommodated within the generalized *LCH* provided that motive satisfies Hypothesis I above—and the limited evidence available appears to support this assumption.

The generalized *LCH* still implies the basic Propositions 1 to 5. On the other hand, Proposition 6 must be released: the generalization of the basic model points to a number of variables that could affect wealth and saving. These include demographic characteristics like the dependency ratio, the rate of return on wealth, household access to credit, and the strength of the bequest motive. Another potentially important variable is Social Security, though its systematic effect on saving has so far proven elusive, a failure not convincingly accounted for by its having two offsetting effects on private saving (cf. Section II, Part C, subsection 3, above).

Allowing for a significant bequest motive raises the issue of its importance. How large a portion of wealth can be traced to this motive, as against true life cycle saving (i.e., hump plus precautionary)? Unfortunately, it seems impossible at present to give a well-founded answer to the question. We know that the share of wealth received through inheritance can be placed at 1/5 to 1/4 for

the United States (and presumably the United Kingdom), but this information is of little help. On the one hand, we know that in a growing economy, if all the inheritance resulted from the bequest motives, the share would tend to *underestimate* its "importance." On the other hand, the observed share is upward biased to the extent that it reflects not just the bequest motive, but also that portion of bequests which arise from the precautionary motive. We do not know how total bequests are split between the two. There is evidence suggesting that the bequest motive is not very important. Thus, in a 1962 survey (Projector and G. Weiss, 1964), only 3 percent of the respondents gave as a reason for saving, "To provide an estate for the family." However, the proportion rises with wealth, reaching 1/3 for the top class (1/2 million 1963 dollars and over). Similar, though somewhat less extreme, results are reported in a Brookings study (R. Barlow et al., 1966). Thus, the bequest motive seems to be limited to the highest economic classes. This hypothesis is supported by the finding of Menchik and David that for (and only for) the top 20 percent, bequests rise proportionately faster than total resources, something which presumably cannot be explained by the precautionary motive. Furthermore, it is consistent, incidentally, with the observation that the decline in wealth with age tends to be more pronounced and systematic in terms of the median than of the mean. But, then the top fifth of the income distribution can be expected to account for substantially more than 1/5 of all bequests. Thus, there is, at present, no basis for estimating, or even placing bounds on, the importance of the bequest motive. My hunch, based on preliminary analysis, is that hump plus precautionary wealth is likely to account for well over half—but this is only conjecture, to be probed by future research.

III. Policy Implications

Limitations of space make it impossible to pursue a systematic analysis of policy issues for which the *LCH* has implications that are significantly different from those derivable by the standard Keynesian consumption function or refinements thereof. I will, however, list some of the major areas of applications with a brief statement of the *LCH* implications:

1. Short-Run Stabilization Policy

(*i*) *The Monetary Mechanism*: The fact that wealth enters importantly in the short-run consumption function means that monetary policy can affect aggregate demand not only through the traditional channel of investment but also through the market value of assets and consumption. (See my 1971 article.)

(*ii*) *Transitory Income Taxes*: Attempts at restraining (or stimulating) demand through transitory income taxes (or rebates) can be expected to have small effects on consumption and to lower (raise) saving because consumption depends on a life resources which are little affected by a transitory tax change (empirically supported). (See the literature cited in my paper with Charles Steindel, 1977, and my paper with Sterling, 1986.)

2. Long-Run Propositions

(*i*) *Consumption Taxes*: A progressive tax on consumption is more equitable than one on current income because it more nearly taxes permanent income (quite apart from its incentive effects on saving.)

(*ii*) *Short and Long-Run Effects of Deficit Financing*: Expenditures financed by deficit tends to be paid by future generations; those financed by taxes are paid by the current generation. The conclusion rests on the proposition that private saving, being controlled by life cycle considerations, should be (nearly) independent of the government budget stance (myself and Sterling), and therefore private wealth should be independent of the national debt (my 1984 paper). It follows that the national debt tends to crowd out an equal amount of private capital at a social cost equal to the return on the lost capital (which is also approximately equal to the government interest bill).

This conclusion stands in sharp contrast to that advocated by the so-called Ricardian

Equivalence Proposition (Robert Barro, 1974) which holds that whenever the government runs a deficit, the private sector will save more in order to offset the unfavorable effect of the deficit on future generations.

Of course, to the extent that the government deficit is used to finance productive investments, then future generations also receive the benefit of the expenditure, and letting them pay for it through deficit financing may be consistent with intergenerational equity.

In an open economy, the investment crowding-out effect may be attenuated through the inflow of foreign capital, attracted by the higher interest that results from the smaller availability of investable funds. However, the burden on future generations is roughly unchanged because of the interest to be paid on the foreign debt.

Finally, if there is slack in the economy, debt-financed government expenditures may not crowd out investment, at least if accompanied by an accommodating monetary policy, but may, instead, raise income and saving. In this case, the deficit is beneficial, as was held by the early Keynesians; however, the debt will have a crowding-out effect once the economy returns to full employment. *LCH* suggests that to avoid this outcome, a good case can be made for a so-called cyclically balanced budget.

REFERENCES

Ackley, Gardner, "The Wealth-Saving Relationship," *Journal of Political Economy*, April 1951, *59*, 154–61.

Ando, A., "The Savings of Japanese Households: A Micro Study Based on Data from the National Survey of Family Income and Expenditure, 1974 and 1979," Economic Planning Agency, Government of Japan, 1985.

______ **and Kennickell, A.,** "How Much (or Little) Life Cycle Is There in Micro Data? Cases of U.S. and Japan," paper presented at Instituto Bancario San Paolo di Torino Conference to Honor Franco Modigliani, in Martha's Vineyard, MA, September 19–21, 1985 (Conference Proceedings forthcoming).

______ **et al.,** "Lags in Fiscal and Monetary Policy," in *Stabilization Policies*, Commission on Money and Credit, Englewood Cliffs: Prentice-Hall, 1963.

Avery, R. B. et al., "Survey of Consumer Finances, 1983: A Second Report," *Federal Reserve Bulletin*, December 1984, *70*, 857–68.

Barlow, R., Brazer, H. E. and Morgan, J. N., *Economic Behavior of the Affluent*, Washington: The Brookings Institution, 1966.

Barro, R. J., "Are Government Bonds Net Wealth?," *Journal of Political Economy*, November/December 1974, *82*, 1095–117.

Bentzel, R., "Nagra Synpunkter pa Sparandets Dynamik," in *Festskrift Tillagnad Halvar Sundberg* (Uppsala Universitetes Arsskrift 1959:9), Uppsala, 1959.

Blinder, A., Gordon, R. A. and Wise, D., "Social Security, Bequests and the Life Cycle Theory of Saving: Cross-Sectional Tests," in F. Modigliani and R. Hemming, eds., *The Determinants of National Saving and Wealth*, New York: St. Martins Press, 1983, 89–122.

Boskin, M. J. and Hurd, M. D., "The Effect of Social Security on Early Retirement," *Journal of Public Economics*, December 1978, *10*, 361–77.

Brady, D. S. and Friedman, R. D., "Savings and the Income Distribution," NBER *Studies in Income and Wealth*, No. 9, New York, 1947.

Brown, T. M., "Habit Persistence and Lags in Consumer Behavior," *Econometrica*, July 1952, *20*, 355–71.

Davies, J. B., "Uncertain Lifetime, Consumption, and Dissaving in Retirement," *Journal of Political Economy*, June 1981, *89*, 561–77.

Diamond, P. A. and Hausman, J. A. "Individual Retirement and Savings Behavior," *Journal of Public Economics*, February-March 1984, *23*, 81–114.

Duesenberry, James S., *Income, Saving and the Theory of Consumer Behavior*, Cambridge: Harvard University Press, 1949.

Feldstein, M., "Social Security, Induced Retirement, and Aggregate Capital Accumulation," *Journal of Political Economy*, September-October 1974, *82*, 905–26.

______, "Social Security and Private Savings: International Evidence in an Extended Life-Cycle Model," in his and R. Inman, eds., *The Economics of Public Services*, London: Macmillan, 1977.

Fisher, F. and Brown, R., "Negro-White Savings Differentials and the Modigliani-Brumberg Hypothesis," *Review of Economics and Statistics*, February 1958, *40*, 79–81.

Fisher, Irving, *The Theory of Interest*, New York: Macmillan, 1930.

Fisher, J., "The Economics of an Aging Population, A Study in Income, Spending and Savings Patterns of Consumer Units in Different Age Groups, 1935-36, 1945 and 1949," unpublished doctoral dissertation, Columbia University, 1950.

Friedman, M., *A Theory of the Consumption Function*, Princeton: Princeton University Press, 1957.

Goldsmith, R. W., *A Study of Saving in the United States*, Princeton: Princeton University Press, 1956.

Hamburger, W., "Consumption and Wealth," unpublished doctoral dissertation, University of Chicago, 1951.

Harrod, R. F., *Towards a Dynamic Economics*, London, 1948.

Houthakker, H. S., "An International Comparison of Personal Saving," *Bulletin of the International Statistical Institute*, 1961, *38*, 55–70.

______, "On Some Determinants of Saving in Developed and Underdeveloped Countries," in E. A. G. Robinson, ed., *Problems in Economic Development*, London: Macmillan, 1965.

Hurd, M. D., "Savings and Bequests," Working Paper No. 1826, NBER, 1986.

Keynes, J. M., *General Theory of Employment, Interest and Money*, New York: Harcourt, Brace, 1936.

Kennickell, A., "An Investigation of Life Cycle Savings Behavior in the United States," unpublished doctoral dissertation, University of Pennsylvania, 1984.

King, M. A. and Dicks-Mireaux, L-D. L., "Asset Holdings and the Life-Cycle," *Economic Journal*, June 1982, *92*, 247–67.

Kotlikoff, L. J., Spivak, A. and Summers, L., "The Adequacy of Savings," *American Economic Review*, December 1982, *72*, 1056–69.

______ **and Summers, L.,** "The Role of Intergenerational Transfers in Aggregate Capital Accumulation," *Journal of Political Economy*, August 1981, *89*, 106–32.

Kuznets, S., *National Income: A Summary of Findings*, NBER, New York: Arno Press, 1946.

Leff, N., "Dependency Rates and Saving Rates," *American Economic Review*, December 1969, *59*, 886–96.

Lydall, Harold, "The Life Cycle in Income, Saving and Asset Ownership," *Econometrica*, April 1955, *23*, 131–50.

Masson, Andre, "A Cohort Analysis of Wealth-Age Profiles Generated by a Simulation Model in France (1949–1975)," *Economic Journal*, March 1986, *96*.

Menchik, P. L. and David, M., "Income Distribution, Lifetime Savings, and Bequests," *American Economic Review*, September 1983, *73*, 672–90.

Mirer, T. W., "The Wealth-Age Relationship among the Aged," *American Economic Review*, June 1979, *69*, 435–43.

Modigliani, F., "Fluctuations in the Saving-Income Ratio: A Problem in Economic Forecasting," in NBER *Studies in Income and Wealth*, No. 11, University Microfilms, 1949.

______, "The Life Cycle Hypothesis of Saving and Inter-Country Differences in the Saving Ratio," in W. A. Eltis et al., eds., *Induction, Growth and Trade*, Essays in Honor of Sir Roy Harrod, London: Clarendon Press, 1970.

______, "Monetary Policy and Consumption: Linkages via Interest Rate and Wealth Effects in the FMP Model," in *Consumer Spending and Monetary Policy: The Linkages*, Conference Series No. 5, Federal Reserve Bank of Boston, Boston, 1971.

______, "The Life Cycle Hypothesis of Saving Twenty Years Later," in M. Parkin, ed., *Contemporary Issues in Economics*, Manchester University Press, 1975.

______, "The Economics of Public Deficits," paper presented at Conference in Memory

of Abba Lerner, Tel Aviv University, Israel, May 28–31, 1984 (Proceedings forthcoming).

______, "Measuring the Contribution of Intergenerational Transfers to Total Wealth: Conceptual Issues and Empirical Findings," paper presented at Modeling the Accumulation and Distribution of Personal Wealth seminar, Paris, France, September 10–11, 1985 (Proceedings forthcoming).

______ **and Ando, A.**, "Tests of the Life Cycle Hypothesis of Savings: Comments and Suggestions," *Bulletin of the Oxford University Institute of Statistics*, 1957, 99–124.

______ **and** ______, "The 'Permanent Income' and the 'Life Cycle' Hypothesis of Saving Behavior: Comparison and Tests," in *Consumption and Saving*, Vol. 2, Wharton School of Finance and Commerce, University of Pennsylvania, 1960.

______ **and** ______, "The 'Life Cycle' Hypothesis of Saving: Aggregate Implications and Tests," *American Economic Review*, March 1963, *53*, 55–84.

______ **and Brumberg, Richard,** "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data," in K. Kurihara, ed., *Post-Keynesian Economics*, New Brunswick: Rutgers University Press, 1954.

______ **and** ______, "Utility Analysis and Aggregate Consumption Function: An Attempt at Integration," in A. Abel, ed., *Collected Papers of Franco Modigliani*, Vol. 2, Cambridge: MIT Press, 1979.

______, **Mason, A. and Sterling, A.**, "Effect of Fiscal Policy on Saving: Evidence from an International Cross-Section," paper in progress.

______ **and Steindel, C.**, "Is a Tax Rebate an Effective Tool for Stabilization Policy?," *Brookings Papers on Economics Activity*, 1:1977, 175–203.

______ **and Sterling, A.**, "Determinants of Private Saving with Special Reference to the Role of Social Security—Cross-Country Tests," in F. Modigliani and R. Hemming, eds., *The Determinants of National Saving and Wealth*, New York: St. Martins Press, 1983, 24–55.

______ **and** ______, "Government Debt, Government Spending, and Private Sector Behavior: A Comment," *American Economic Review*, forthcoming 1986.

Munnell, A., *The Effect of Social Security on Personal Saving*, Cambridge: Ballinger Press, 1974.

Neisser, H. P., "The Economics of a Stationary Population," *Social Research*, November 1944, *11*, 470–90.

Projector, D., *Survey of Changes in Family Finances*, Washington: Board of Governors of the Federal Reserve System, 1968.

______ **and Weiss, G.**, *Survey of Financial Characteristics of Consumers*, Washington: The Board of Governors of the Federal Reserve, 1964.

Reid, M. G., "The Relation of the Within-Group Permanent Component of Income to the Income Elasticity of Expenditure," unpublished paper.

Ricci, U., "L'offerta del Risparmio," Part I, *Giornale degli Economisti*, 1926; Part II, ibid, 1926; "Ancora Sull'Offerta del Risparmio," ibid, 1927.

Shefrin, H. M. and Thaler, R., "Life Cycle Vs. Self-Control Theories of Saving: A Look at the Evidence," unpublished paper, 1985.

Shiba, Tsunemasa, "The Personal Savings Functions of Urban Worker Household in Japan," *Review of Economics and Statistics*, May 1979, *61*, 206–13.

Shorrocks, A. F., "The Age-Wealth Relationship: A Cross-Section and Cohort Analysis," *Review of Economics and Statistics*, May 1975, *57*, 155–63.

Tobin, James, "Life Cycle Saving and Balanced Growth," in W. Fellner et al., eds., *Ten Economic Studies in the Tradition of Irving Fisher*, New York: Wiley & Sons, 1967.

Royal Commission on the Distribution of Income and Wealth, *Report No. 5.*, *Third Report on the Standing Reference*, London: HMSO, 1977.

Is the Stabilization of the Postwar Economy a Figment of the Data?

By CHRISTINA D. ROMER*

One of the most recurrent empirical generalizations about the U.S. economy is that the prewar economy was substantially more volatile than the postwar economy. It is widely accepted that the business cycle before World War II (or before World War I for that matter) was decidedly more severe than the cycle after 1947. The source of this belief is simply every conventional indicator of macroeconomic performance; industrial production, unemployment, and Gross National Product all show larger cyclical fluctuations in the late 1800's and early 1900's than after World War II.

This paper challenges part of the stylized fact that the prewar economy was substantially more volatile than the postwar economy. It provides an examination of the conventional industrial production series for the pre-World War I and post-World War II periods and shows that the apparent stabilization of this series is actually a figment of the data. I find that the methods used to construct the historical series exaggerate cyclical fluctuations in industrial production. When this exaggeration is taken into account, there is very little stabilization between the pre-1914 and the post-1947 eras.

By itself, this study of the historical industrial production data challenges some of the existing empirical studies of the stabilization of the postwar economy. For example, a recent paper by J. Bradford DeLong and Lawrence Summers (1984) uses the conventional industrial production series and real Gross National Product series to argue that cycles have become much less severe over time. By showing that the industrial production series has not stabilized over time, the present study undermines part of the empirical regularity DeLong and Summers seek to explain.

This study of the historical industrial production data is also part of a larger project. In two other papers (1985 and 1986), I examine the historical unemployment and Gross National Product data. These two studies yield results very similar to those for industrial production. In all three cases there exist fundamental inconsistencies between the historical and modern series that account for much of the damping of cyclical fluctuations between the prewar and postwar eras.

In conjunction with these other studies of historical macroeconomic data, this study of the prewar industrial production series challenges another strain of the stabilization literature. To many, what is most striking about the twentieth-century business cycle is not that particular indicators have stabilized, but rather that nearly all macroeconomic series show less severe fluctuations in the postwar era (see, for example, Arthur Burns, 1960, and Robert Lucas, 1977). The fact that the present study finds that the historical industrial production data are excessively volatile in the same way that the historical unemployment and *GNP* data are refutes this finding. Errors in the three series rather than genuine economic changes account for the apparent stabilization of these key macroeconomic series.

Historical Industrial Production Data. The historical industrial production series that I examine in this paper is that constructed by Edwin Frickey in his book *Production in the United States, 1860–1914.* Frickey's index is one of the least recognized but most often used macroeconomic series. Frickey's index is traditionally paired with the Federal Reserve Board (FRB) index of industrial

*Woodrow Wilson School, Princeton University, Princeton, New Jersey 08544. I thank Olivier Blanchard, Rudiger Dornbusch, David Romer, Robert Solow, Peter Temin, and an anonymous referee for extremely helpful comments and suggestions.

production in manufacturing which begins in 1919 to form a series on manufacturing production going back to 1860 (see, for example, *Historical Statistics of the United States*, 1975). Frickey's index is also often combined with other historical series on mining and utilities production to form a historical extension of the total FRB index of industrial production (see, for example, G. Warren Nutter, 1962). In addition to being the key historical industrial production series, Frickey's index is also the basis for other conventional output series. For example, John Kendrick's (1961) historical estimates of total output in manufacturing are formed by using Frickey's series to interpolate between census-year benchmarks. Thus, all annual movements in this important output series are derived directly from Frickey's industrial production series.

Frickey's index is also important because it uses essentially the same methodology and includes many of the same commodities as do several other prewar indexes of industrial production. (See, for example, Frederick Mills, 1932, Warren Persons, 1931, and Walter Stewart, 1921.) As a result, Frickey's index is representative of a class of output measures. Hence, any errors found in the Frickey index will almost certainly be present in these other series.

Because of its widespread use, Frickey's index has affected many of our views about the U.S. economy before World War I. When compared to the modern FRB index of industrial production, Frickey's index is substantially more volatile. For example, the average peak-to-trough change in Frickey's index for 1866–1914 is 26 percent greater than that of the modern FRB index of industrial production in manufacturing for 1947–82. Thus, the large cyclical swings in Frickey's industrial production series have helped generate the belief that the prewar economy was much less stable than the economy after 1947.

Although Frickey's series is often used as if it were the prewar extension of the FRB index of industrial production in manufacturing, the prewar and postwar data are not consistent. Frickey's index is based on a much smaller sample of commodities than is the modern FRB manufacturing index. Furthermore, the types of goods included in Frickey's index are qualitatively different from those included in the FRB index. Whereas the FRB manufacturing index includes data on both materials and finished goods, Frickey's index is based almost entirely on materials and very basic manufactured commodities.

Overview. To see if these differences between the prewar and postwar industrial production data can explain the apparent stabilization of the postwar economy, one must separate true economic changes from the tremendous improvements in data collection procedures. To do this, I construct an exact replication of Frickey's prewar index for the post–1947 period. Replicating Frickey's procedures for the modern period yields an industrial production series that is consistent over time. The pre–1914 and the post–1947 series can be compared to see what, if any, changes have occurred in the economy. Furthermore, the modern Frickey replication can be contrasted with the modern FRB manufacturing index to show the effects of changes in data collection and transformation techniques.

In addition to examining the behavior of an exact postwar replication of Frickey's series, I also consider an updated replication of Frickey's series. The FRB index of materials production provides a modern index that is qualitatively similar to Frickey's series, but is less anachronistic than the exact replication. The FRB materials index can again be compared to the prewar Frickey index to see if consistent industrial production data show any stabilization between the prewar and postwar eras. Analysis of the FRB materials index can also be used to suggest the source of errors in the postwar exact replication of Frickey's index.

Once the source of the errors in the postwar replications of Frickey's index has been identified, it is then possible to see if the same source of errors exists in the prewar era. To preview, I find that the characteristics of the economy that cause Frickey's methods to exaggerate cyclical movements in the postwar era have not changed over time.

Thus, it is likely that Frickey's prewar index is excessively volatile.

The various steps in this analysis of the historical industrial production series are organized as follows. Section I describes Frickey's index and discusses the creation of a consistent postwar series. Section II presents a detailed comparison of the pre-1914 Frickey series and both the exact postwar replication of Frickey's series and the postwar FRB materials index. Similar business cycle analytics are used to compare these series to the postwar FRB index of industrial production in manufacturing. Section III examines why the historical methods yield a postwar series that is excessively volatile. Section IV presents evidence that the historical methods generate similar errors in the prewar era. Finally, Section V compares the results of this study of the industrial production data to those of my other studies of the unemployment and *GNP* data. It also discusses the possible implications of the findings for the effectiveness of stabilization policy.

I. Replicating Frickey's Procedures for the Postwar Period

The procedures Frickey uses to construct a prewar index of industrial production are very similar to those the Federal Reserve Board uses today. Both indexes are formed by combining data on the physical quantity of various manufactured goods. Both use value-added weights to combine numerous individual indexes of production into a single index of industrial output. Although there are minor differences in the classification of various products and in benchmarking procedures, the only major difference between the two series is the number and range of commodities included in each index. The FRB manufacturing index includes over 200 commodities; Frickey's index includes 40 commodities.

The series included in Frickey's index are not only fewer in number, but also qualitatively different from those included in the modern FRB manufacturing index. The quantity data available for the turn of the century cover only very basic commodities. Production figures exist primarily for materials and for goods early in the manufacturing process.[1] For example, there exist good data on pig iron production but no estimates of the production of tools and machinery. Similarly, there are figures on the amount of lumber produced but none on the production of flooring or other millwork products. Out of necessity, Frickey uses these available data on the production of materials to proxy for the output of more fabricated products.

Whenever Frickey uses materials to proxy for output, he tries to convert this data on materials produced to an estimate of materials consumed. That is, he attempts to estimate the amount of materials that are actually used in the domestic production of finished goods. In general, to estimate materials consumed, Frickey only corrects the existing materials produced series for fluctuations in foreign trade. In most cases the size of these corrections are so small that the production and consumption series are nearly indistinguishable.

To analyze possible errors in Frickey's prewar index, I examine two postwar replications of Frickey's series. The first is an exact replication of Frickey's methods. I form a postwar series using methods and a sample of commodities that are identical to those Frickey uses. The resulting series is consistent with the prewar Frickey index in the naive sense that the two series are formed in exactly the same way.

The second postwar extension that I consider is an updated replication of Frickey's methods. This replication tries to preserve the essential elements of Frickey's methods while taking into account the many changes that have occurred in the economy. This replication seeks to hold constant the relationship between the index and the underlying economy. That is, this updated replication is designed to preserve the limitations of Frickey's original index, but to remove the additional errors that result from replicating

[1] There are also data on goods that were taxed; specifically alcoholic beverages and tobacco products. While Frickey does include these data in his index of industrial production, these series receive very little weight in the total index.

exactly Frickey's prewar methods and sample of commodities for the larger, more advanced postwar economy.

A. *Exact Replication*

Forming an exact postwar replication of Frickey's index is fairly straightforward. The 40 commodities that Frickey includes in his index represent most of the important, high-growth materials of the prewar era. Among the series included in Frickey's index are wheat flour produced, lumber produced, pig iron consumed, and petroleum produced. Data on nearly all the goods represented in Frickey's index are still collected today. Most of the modern series can be found in either *Historical Statistics of the United States* or *Business Statistics*, the biennial supplement to the *Survey of Current Business*.

I follow Frickey very closely in combining the 40 modern series into an index of industrial production. To combine the various commodity quantity series, Frickey first converts each series into an index based in 1899. These indexes are then combined by taking a weighted average of the individual indexes. For Frickey's original index, these weights are based on the percent of total value-added that each commodity accounted for in 1899. To replicate Frickey's procedures, I choose 1967 as the base year. The weights are derived from data on value-added from the 1967 *Census of Manufactures*.

The actual derivation of the weights is complicated because the allocation of value-added to various commodities is done on the basis of all of the products for which the commodities are proxying. For example, paper consumed is weighted according to the entire value-added in the output of the printing and publishing industry. Similarly, pig iron consumed is allocated the value-added of all iron and steel "end products" except for the few for which output series exist. Fortunately, Frickey is detailed enough that it is possible to assign weights as he does.

The results of applying Frickey's procedures in the postwar era are shown in Table 1. The table shows Frickey's original index for 1866–1914 and my exact replication of his index for 1947–82.

B. *Updated Replication*

In addition to the exact replication, it is also desirable to examine an updated extension of Frickey's index. The FRB index of industrial production of materials is a readily available example of such an updated replication. The FRB materials index preserves the reliance on materials that is obvious in Frickey's index, but greatly expands the sample of materials included in the index. In addition to including most of the goods represented in Frickey's index, the FRB materials index also contains most manufactured materials such as plastics and synthetic rubber. Because the FRB materials index includes most of the important, high-growth materials, it measures the trend of industrial production in the postwar era more accurately than does the anachronistic exact Frickey replication.

The FRB materials index is a good updated replication of Frickey's index because it takes into account the changing degree of fabrication in the economy. The materials included in the FRB materials index are in general somewhat more fabricated than those in Frickey's index. Goods such as engine parts and automobile windshields are classified as materials in the FRB index. This upgrading of the commodities included in the materials index compensates for the increasing fabrication of finished goods in the United States.

While the FRB materials index is a convenient update of Frickey's original index, it is in some sense too good an index. The FRB materials index includes a much larger sample of goods and many more new commodities than does the prewar Frickey index. Thus, it is likely that the FRB materials index does not preserve all the limitations of Frickey's prewar index.[2]

[2]The only way in which the FRB materials index may be a slightly less accurate indicator of total production than is Frickey's original index is that the FRB materials index makes no correction for foreign trade. However, the effect of this particular correction is very small because the volume of trade in the materials included in Frickey's index is minimal.

TABLE 1—INDEX OF INDUSTRIAL PRODUCTION

Year	Frickey (1866–1914)	Year	Frickey Replication (1947–82)	FRB Materials (1947–82)
1866	21	1947	76.78	39.5
1867	22	1948	78.36	41.2
1868	23	1949	70.83	37.6
1869	25	1950	82.19	45.0
1870	25	1951	84.29	49.8
1871	26	1952	78.03	50.5
1872	31	1953	85.42	56.1
1873	30	1954	74.92	51.8
1874	29	1955	88.25	61.3
1875	28	1956	86.47	62.8
1876	28	1957	85.84	62.8
1877	30	1958	74.22	56.5
1878	32	1959	82.10	65.2
1879	36	1960	83.80	66.1
1880	42	1961	82.28	66.2
1881	46	1962	86.01	72.1
1882	49	1963	89.77	76.7
1883	50	1964	97.40	82.9
1884	47	1965	101.91	92.4
1885	47	1966	104.48	100.7
1886	57	1967	100.00	100.0
1887	60	1968	104.47	106.5
1888	62	1969	106.07	112.5
1889	66	1970	102.01	109.2
1890	71	1971	101.78	111.3
1891	73	1972	107.02	122.3
1892	79	1973	113.35	133.9
1893	70	1974	106.72	132.4
1894	68	1975	96.06	115.5
1895	81	1976	105.25	131.7
1896	74	1977	104.86	138.6
1897	80	1978	110.21	148.3
1898	91	1979	109.91	156.4
1899	100	1981	96.86	147.6
1900	100	1981	104.45	151.6
1901	111	1982	82.27	133.7
1902	127			
1903	126			
1904	121			
1905	140			
1906	152			
1907	156			
1908	127			
1909	166			
1910	172			
1911	162			
1912	194			
1913	203			
1914	192			

II. Comparing the Pre-1914 and the Post-1947 Indexes of Industrial Production

Having created two versions of a consistent industrial production series, it is possible to see how much of the apparent stabilization of the postwar economy is due to improvements in the data. By comparing Frickey's prewar index to either its exact or updated postwar replication, it is possible to see what the stylized facts about the economy would have been in the absence of the Federal Reserve Board index of industrial production in manufacturing. It is also useful to contrast both the postwar extensions of Frickey's index with the actual postwar FRB manufacturing index. This will help to identify the magnitude and the direction of the errors caused by using the historical methods to create postwar data.

Given that the postwar series were created using Frickey's methods, the only valid comparisons are between the prewar Frickey data and the postwar series. Since the Frickey series only exists from 1866–1914,[3] the periods of comparison must be the pre-World War I era and the post-World War II era. These are, however, both appropriate and natural eras to consider. The years between the close of the Civil War and the beginning of World War I cover a period that is traditionally thought to be very volatile. If there has indeed been a genuine damping of business cycle fluctuations over time, then the pre-1914 economy should certainly be more volatile than the economy after 1947 when consistent data are compared. Furthermore, the pre-1914 era is a period in which government monetary and fiscal policy is generally considered to have been of reasonably little importance in comparison to the importance of these policies in the postwar era. Hence a comparison of the two periods can be considered a comparison of the prepolicy and postpolicy eras.

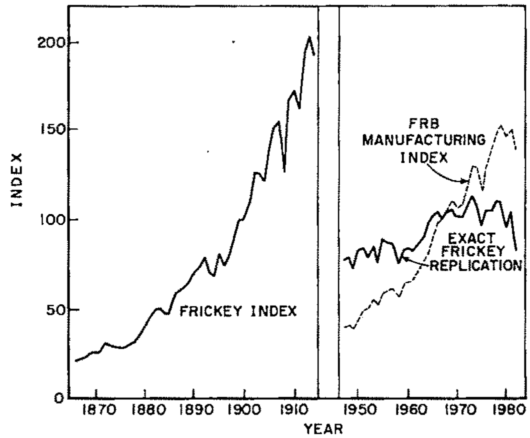


FIGURE 1. INDUSTRIAL PRODUCTION, 1866–1914 AND 1947–82

A. *Trends*

Figure 1 shows Frickey's original series for 1866–1914 and the exact postwar replication for 1947–82. It also shows the Federal Reserve Board index of industrial production in manufacturing for 1947–82.[4]

In Figure 1 the most noticeable difference between the various series is that the trend of the exact postwar replication of Frickey's series is much flatter than the trend of either Frickey's prewar series or the postwar FRB manufacturing index. The average growth rate of Frickey's series for 1866–1914 is 4.6 percent per year and that of the FRB manufacturing index for 1947–82 is 3.6 percent. However, the exact Frickey replication for the postwar period shows almost no growth between 1947 and 1982. The average growth rate of this series is only 0.2 percent per year.

This result is clearly due to the fact that the exact postwar replication is based on a very out-of-date sample of commodities. Though not shown in Figure 1, the updated Frickey replication (the FRB materials index) shows roughly the same growth as the

[3] Frickey actually forms data for 1860–65 as well. However, he suggests that the data before 1866 are of substantially lower quality than the data after 1866 (p. 3).

[4] The FRB index of industrial production in manufacturing comprises 87 percent of the total FRB industrial production index. It excludes production in mining and utilities.

TABLE 2—MEASURES OF VOLATILITY

Measure	Frickey (1866–1914)	Exact Replication (1947–82)	FRB Materials (1947–82)	FRB Manufacturing (1947–82)
Mean Cyclical Amplitude of Detrended Series[a]	.1398	.1319	.1342	.1081
Standard Deviation of Growth Rates	.0884	.0862	.0797	.0643
Standard Deviation of Deviations from Trend	.0830	.0762	.0728	.0636

[a]Trend industrial production is calculated as the fitted value of a regression of the log of industrial production on a constant and a quadratic trend.

FRB manufacturing index. The average growth rate of the materials index for 1947–82 is 3.5 percent per year. The large discrepancy between the trends of the two postwar Frickey replications shows that the accuracy with which the historical methods measure the level of industrial production depends crucially on whether one replicates the historical procedures exactly or modifies them to include modern goods.

B. *Volatility*

A second noticeable difference among the three series shown in Figure 1 is that the exact postwar replication of Frickey's series is much more volatile than the FRB manufacturing index. While the peaks and troughs of the two series are roughly coincident, the severity of cyclical swings is greater in the exact Frickey replication. This same pattern holds for the FRB materials index as well. It too is substantially more volatile than the postwar FRB manufacturing index. The greater cyclical volatility of both the consistent postwar extensions of Frickey's index makes business cycle fluctuations of these series resemble those of the prewar Frickey index quite closely.

1. *Mean Cyclical Amplitude*. The differences in volatility between all four series can be described and quantified in a variety of ways. Table 2 presents three common measures of volatility. The first of these measures is the mean cyclical amplitude of each series. This measure shows the average percentage fall in industrial production between peaks and troughs of the business cycle. For the four series on industrial production under consideration, the measurement of cyclical amplitudes is complicated by the differences in the trends of the various indexes. To account for this, the cyclical amplitude is calculated as the peak-to-trough change in the logarithm of the detrended index of industrial production. The trend values of each index are estimated as the fitted value of a regression of the logarithm of the index on a constant and quadratic trend.[5] For the calculations in Table 2, the peaks and troughs are defined to be the actual turning points in each detrended series.

From the statistics in Table 2 it is clear that the exact postwar Frickey replication and the FRB materials index are both substantially more volatile than the postwar FRB manufacturing index. For example, the mean cyclical amplitude of the FRB materials index is approximately 22 percent greater than that of the FRB manufacturing index. At the same time, both the postwar replications of Frickey's series are less volatile than the prewar Frickey series. However, the implied stabilization is very slight. The exact replication is only 6 percent less volatile than the prewar Frickey index and the FRB materials index is only 4 percent less volatile than Frickey's series. Thus, a comparison of consistent data

[5]The quadratic trend was chosen because it appears to fit all four series substantially better than a linear trend.

TABLE 3—BUSINESS CYCLES

Frickey (1866–1914)		Exact Replication (1947–82)		FRB Materials (1947–82)		FRB Manufacturing (1947–82)	
Peak-Trough	Percent Decline[a]	Peak-Trough	Percent Decline[a]	Peak-Trough	Percent Decline[a]	Peak-Trough	Percent Decline[a]
1869–71	.0556	1947–49	.1237	1947–49	.1581	1947–49	.1219
1872–76	.2917	1951–52	.0963	1951–52	.0376	1951–52	.0093
1882–85	.1845	1953–54	.1488	1953–54	.1297	1953–54	.1177
1887–88	.0149	1955–58	.2199	1955–58	.2240	1955–58	.1597
1890–91	.0199	1960–61	.0311	1959–61	.0740	1959–61	.0664
1892–94	.2454	1966–67	.0524	1966–67	.0463	1966–67	.0184
1895–96	.1382	1969–71	.0536	1969–71	.0836	1969–70	.1001
1899–1900	.0478	1973–75	.1723	1973–75	.2141	1973–75	.1791
1902–04	.1441	1978–82	.2891	1979–82	.2403	1979–82	.2008
1906–08	.2755						
1909–11	.1203						

[a]The percentage decline is measured as the difference between the logarithms of the peaks and troughs of the detrended series.

does not reveal the dramatic damping of business cycle fluctuations apparent in the inconsistent series.

The larger cyclical amplitude of the postwar Frickey replications is important because it shows that Frickey's methods systematically exaggerate cyclical fluctuations in the postwar period. For most cycles both the exact postwar Frickey replication and the FRB materials index show a larger percentage fall in output than does the postwar FRB manufacturing index. This can be seen in Table 3 which shows the peak-to-trough declines in industrial production for each cycle for these three indexes. The historical methods clearly overstate cyclical movements rather than merely add noise to the series.

2. *Standard Deviation of Growth Rates*. The standard deviation of the growth rate of industrial production is another measure of volatility. It measures how much the change in output varies from year to year. The statistics in Table 2 again show that the growth rate of industrial production is much more variable for either of the postwar Frickey replications than for the modern FRB manufacturing series. The growth rate of the exact Frickey replication for 1947–82 is approximately 29 percent more variable than the FRB manufacturing series for the same time period. At the same time, both the postwar extensions of Frickey's series are somewhat less volatile than Frickey's prewar index. Thus, for this measure of volatility, consistent data indicate that industrial production may have stabilized some, but not nearly as much as a comparison of the Frickey and FRB manufacturing data would suggest.

3. *Standard Deviation of Deviations from Trend*. The standard deviation of the deviations of industrial production from trend provides a final measure of the volatility of each series. This measure indicates the variability of yearly cyclical movements. In Table 2, I report estimates using a quadratic trend.[6] The standard deviations indicate that both the postwar Frickey replications are more volatile than the postwar FRB manufacturing series, but slightly less volatile than the pre-1914 Frickey index. For example, the standard deviation of deviations from trend of the exact postwar replication of Frickey's index is 18 percent greater than that of the postwar FRB manufacturing index and 9

[6]When a linear trend is used to detrend industrial production, the standard deviations of deviations from trend are: Frickey (1866–1914) = .0830; Exact Frickey Replication (1947–82) = .0835; and FRB Materials (1947–82) = .0831; FRB Manufacturing (1947–82) = .0714.

percent smaller than that of the prewar Frickey series.

Despite some differences between the various measures of volatility, the results of these three comparisons all point to a similar conclusion: a substantial amount of the apparent stabilization of the postwar index of industrial production is due to improvements in the data. Depending on which series and measure are used, somewhere between half and all of the observed stabilization is the result of comparing inconsistent data. When a consistent series is compared over time, the amplitude of the cycle is roughly similar before World War I and after World War II. Furthermore, while growth rates and the deviations of industrial production from trend have stabilized some, the change over the twentieth century has been mild, not dramatic.

C. *Significance Tests*

One question raised by all the measures of volatility is whether the differences between various indexes are statistically significant. From the perspective of this paper, however, significance is not a major issue. The existence of the stylized fact that the economy has stabilized implies a general consensus that the difference in volatility between the pre- and postwar series is important. If this difference is not statistically significant, then the stylized fact is on shaky ground regardless of inconsistencies in the data. If the difference is significant, then the comparison of the postwar Frickey replications and FRB manufacturing index provides an estimate of how much of the difference arises from changes in data collection procedures.

If one is willing to make distributional assumptions, it is nevertheless possible to test whether various differences in volatility are significantly different from zero. Such significance tests are straightforward in the case of differences in mean cyclical amplitudes. For comparison of the prewar and postwar series, the usual test of the difference between two means can be used. This test assumes that the two samples are independent random samples from populations that are distributed normally with the same variance. For comparison of various postwar series, it is necessary to use a paired *t*-test because the two samples are clearly not independent.

The test for whether differences in the standard deviations of the growth rates of various series are significant is also straightforward. Under the assumptions of normality and independence, the ratio of the two variances of growth rates is distributed F with degrees of freedom corresponding to the size of the two samples.[7] The same test can be used to compare the standard deviations of deviations from trend.

The test statistics for the various significance tests are shown in Table 4. One result is that using the traditional inconsistent data, the apparent stabilization of the postwar economy is significant. When Frickey's prewar data are compared to the modern FRB index of industrial production in manufacturing, it is generally possible to reject the hypothesis that the volatility of the two series is the same.

A second result is that the slight stabilization shown in the consistent data is not significant. For all three measures of volatility, it is not possible to reject the hypothesis that the prewar Frickey index and both the postwar extensions of Frickey's index are equally volatile. Furthermore, it is also not possible to reject the hypothesis that the exact Exact Replication replication and the postwar FRB materials index are equally volatile. Thus, it is reasonable to conclude that Frickey's prewar index and both the possible consistent postwar extensions have equally severe cyclical movements.

D. *Length and Timing of Cycles*

The fact that Frickey's methods overstate cyclical movements in the postwar period has important implications for the timing and duration of cycles. In several cases this

[7] For growth rates the assumption of independence is probably reasonable. Because the logarithms of various indexes of industrial production are fairly close to random walks, the growth rates are nearly serially uncorrelated. For deviations from trend, the assumption of independence is clearly much less realistic.

TABLE 4—SIGNIFICANCE TESTS

Differences in Mean Cyclical Amplitude[a]	
Mean (Frickey)-Mean (Exact Replication)	.0079
	(.1868)
Mean (Frickey)-Mean (FRB Materials)	.0056
	(.1362)
Mean (Frickey)-Mean (FRB Manufacturing)	.0317[b]
	(.8085)
Mean (Exact Replication)-Mean (RFB Materials)	−.0023
	(−.1783)
Mean (Exact Replication)-Mean (FRB Manufacturing)	.0238[c]
	(1.444)
Mean (FRB Materials)-Mean (FRB Manufacturing)	.0260[e]
	(3.416)
Equality of Standard Deviations of Growth Rates	
SD^2 (Frickey)/SD^2 (Exact Replication)	$F=1.052$
SD^2 (Frickey)/SD^2 (FRB Materials)	$F=1.230$
SD^2 (Frickey)/SD^2 (FRB Manufacturing)	$F=1.890$[d]
Equality of Standard Deviations of Deviations from Trend	
SD^2 (Frickey)/SD^2 (Exact Replication)	$F=1.186$
SD^2 (Frickey)/SD^2 (FRB Materials)	$F=1.300$
SD^2 (Frickey)/SD^2 (FRB Manufacturing)	$F=1.703$[d]

[a]The t-statistics are shown in parentheses.
[b]Significant at the 80 percent confidence level.
[c]Significant at the 90 percent confidence level.
[d]Significant at the 95 percent confidence level.
[e]Significant at the 99 percent confidence level.

exaggeration turns periods of stagnation into what appear to be genuine recessions. Periods of no growth or very slight downturn in the FRB manufacturing index turn into periods of substantial drops in output in the exact or updated postwar Frickey indexes. This fact is easily seen in Table 3, which shows the peaks and troughs of the exact Frickey replication, the FRB materials index, and the FRB manufacturing index. The peaks and troughs are not necessarily NBER reference cycles. Rather, they correspond to actual highs and lows of the particular detrended annual series.

From the peak-to-trough changes in the three detrended series it is possible to see that the postwar Frickey replications have cycles not present in the FRB manufacturing index. The years 1952 and 1967 appear to be years of significant recession in both the exact Frickey replication and the FRB materials index, while they are only the mildest downturns in the postwar FRB manufacturing index. Because the downturns in the FRB manufacturing index are so slight, it is fair to say that the 1951–52 and 1966–67 cycles do not appear in the true data. On the other hand, these same recessions clearly are genuine cycles in the postwar extensions of Frickey's data.[8]

The presence of additional cycles in the Frickey replications alters one's view of the postwar economy. Rather than looking like a period of long, protracted cycles, the post-1947 era looks more like an era of short, choppy cycles. This is seen when one compares the average duration of cycles. In the FRB index of manufacturing production for 1947–82, a cycle lasts on average 5.3 years. Both the postwar extensions of Frickey's index have cycles substantially shorter than those in the postwar FRB manufacturing series. For the exact replication, the average cycle lasts 3.9 years. For the FRB materials index, the average cycle lasts 4.0 years.

[8]This assertion can be codified by saying that cycles with a percent decline of less than 2 percent should not be counted as genuine cycles.

TABLE 5—SAMPLE AUTOCORRELATIONS

Lags	Frickey (1866–1914)	Exact Replication (1947–82)	FRB Materials (1947–82)	FRB Manufacturing (1947–82)
Percentage Changes				
1	−.195	−.382	−.230	−.162
2	−.304	.122	−.044	−.092
3	.220	−.053	−.056	−.038
4	−.011	−.022	.014	.000
5	−.256	.178	.100	.097
6	−.092	−.178	−.146	−.114
Deviations from Trend				
1	.439	.260	.351	.437
2	.112	.260	.125	.173
3	.124	.087	.023	.065
4	−.135	.082	.041	.051
5	−.360	.108	.014	.020
6	−.297	.097	−.143	−.126

If one makes the same calculation for the prewar Frickey index, the average cycle is longer than that for the postwar Frickey replications. Excluding very small cycles (those with a decline in industrial production of less than 2 percent), the average cycle in the prewar Frickey index lasts 5.0 years. This suggests that when consistent data are compared, cycles are approximately one year shorter in the prewar era than in the postwar era.

Very similar results emerge from a consideration of the autocorrelation functions for each index of industrial production. The first six sample autocorrelations of the percentage changes and the deviations from trend of series are given in Table 5. Although the magnitude of the differences in the autocorrelation functions of the four series is reasonably small, the direction of differences is suggestive. Using percentage changes, the first-order sample autocorrelations of the prewar Frickey index and the postwar FRB manufacturing index are smaller negative numbers than are the first-order sample autocorrelaions of either of the postwar replications of Frickey's index. This suggests that the postwar exact Frickey replication and the FRB materials index exhibit choppier movements than do either the prewar Frickey index or the postwar FRB manufacturing index. Using deviations from trend, the first-order sample autocorrelations of the prewar Frickey index and the postwar FRB manufacturing index are larger positive numbers than are the first-order sample autocorrelations of either of the postwar Frickey replications. This finding is consistent with the view that cycles are more protracted in the prewar Frickey index and the postwar FRB manufacturing index than in the exact postwar Frickey replication or the FRB materials index.

The results of both the simple calculation of the length of cycles and the estimation of sample autocorrelations challenge the traditional view that the length and timing of cycles have been stable between the prewar and postwar eras.[9] While this traditional view is evident in a comparison of the prewar Frickey and the postwar FRB indexes of manufacturing production, it is much less apparent when consistent data are compared over time. When consistent data are examined, cycles in the postwar era appear to be somewhat shorter and less protracted than cycles in the period 1866–1914. This result,

[9]The conventional stylized fact is stated most succinctly in Victor Zarnowitz and Geoffrey Moore. Using the conventional NBER business cycle chronology they conclude "With regard to the total cycle durations, neither the means nor the standard deviations indicate any significant trends. Expansions lengthened and contractions shortened drastically but cycle lengths remain about the same" (1984, p. 7).

combined with the earlier results on cyclical amplitude, may suggest that while policy or institutional changes in the economy have not led to a dramatic decline in the severity of cycles over time, they have led to a noticeable shortening of cyclical fluctuations.

III. The Source of Excess Volatility

Considering the significant differences between the postwar FRB manufacturing index and both postwar replications of Frickey's index, it is important to discover what is causing the postwar replications of Frickey's index to have much larger and more frequent cyclical fluctuations than the FRB index of manufacturing production. Specifically, to be able to argue that the prewar Frickey series is excessively volatile, it is necessary to know where the historical methods go wrong. Only by identifying the source of systematic errors in the postwar constructed series is it possible to see if the same sources exist in the historical period.

Since the methods used to construct both the Frickey and the FRB indexes are very similar, the source of the differences in volatility must lie in the vast differences in the sample of commodities included in the two indexes. The three main discrepancies between the two samples are that Frickey's sample is much smaller, much more biased toward materials and primary goods, and comprised of more outdated goods than is the modern FRB manufacturing index.

A. *Reliance on Materials*

From the comparisons of Section II we already have a great deal of evidence concerning which of these differences is most important. The behavior of the exact postwar Frickey replication shows the combined result of all three discrepancies. By using exactly the same sample of commodities that Frickey uses, the exact replication has all the errors stemming from using the production of a small sample of outdated materials to estimate industrial production. The behavior of the FRB materials index, on the other hand, shows the result of only one discrepancy: the reliance on materials. Because the FRB materials index includes a large sample of modern commodities, it should be free of the problems related to using a small sample of anachronistic commodities.

The statistics in Table 2 show that the mean cyclical amplitudes of the exact Frickey replication and the FRB materials index differ by less than 5 percent. A paired t-test shows that the two amplitudes are not significantly different from one another. The standard deviations of percentage changes and deviations from trend are also very similar for the exact Frickey replication and the FRB materials index. This suggests that despite the large differences in the number of commodities and the number of modern products included in the two indexes, the two postwar extensions of Frickey's index appear to accentuate cyclical movements to nearly the same degree. Since the reliance on materials is the one characteristic that the two indexes have in common, it is likely that the reliance on materials is the key source of the exaggeration of cyclical movements in the postwar extensions of Frickey's index.

B. *Cyclical Movements in Materials Inventories*

It is natural to question why materials are more volatile than manufactured goods in general. A partial explanation is that investment in materials inventories is very procyclical. In a recession, inventories of manufactured materials and supplies are run down tremendously, while in a boom they are increased greatly. This implies that, for a given level of demand, the movement in the production of materials is substantially greater than the actual movements in the consumption of these goods. Thus, even if the consumption of materials were proportional to the output of more fabricated products, the production of materials would show larger cyclical fluctuations than manufactures in general.

To see this more clearly it is useful to examine two identities. Suppose that industrial production is directly proportional to materials consumed. Then

$$(1) \qquad Y_t = YMC_t$$

where Y_t is total output and YMC_t is materials consumed. By definition,

$$(2) \qquad YMC_t = YM_t - \Delta N_t$$

where YM_t is the production of materials and ΔN_t is the change in materials inventories over year t. If materials inventory investment is procyclical, then cyclical movements in materials consumed are always smaller than cyclical movements in materials produced. Empirically, the correlation between changes in materials inventories and the deviations of industrial production from trend is approximately 0.6.[10] Thus, materials inventories movements may explain why a materials index is more volatile than an index of total output.

Given this possible role for inventories, it is useful to see if the size and timing of movements in materials inventories are such that they can explain the observed difference in the cyclical volatility of materials production and the production of all manufactured commodities taken together. One way to see how important inventories are is to see how much correcting the FRB materials index for movements in inventories changes the volatility of the materials index.

To change the materials index from a production to a consumption index, one must first develop a series on the real value of materials inventories. The Bureau of the Census has collected data on materials and supplies inventories since 1954. This is a nominal series that values inventories at the smaller of cost or price. Thus, any form of deflation is inaccurate. Accepting this problem, a reasonable way of deflating is simply to divide the inventory series by the Producer Price Index for intermediate goods. While still inaccurate, this price index does measure many of the relevant price changes.

[10]The change in materials inventories is measured using the Bureau of the Census series on the nominal value of materials and supplies inventories, deflated by the Producer Price Index for intermediate goods. Industrial production is measured using the FRB manufacturing index.

To combine the changes in real inventories with the FRB index of materials production requires a further assumption. The change in real inventories can be represented as

$$(3) \qquad \Delta N_t = \sum_i p_{i,1967} \Delta Q_{i,t}$$

where ΔN_t is the change in the level of real materials inventories in year t, $p_{i,1967}$ is the price of material i in 1967, and $\Delta Q_{i,t}$ is the change in the quantity of inventories of material i in year t. The index of the industrial production of materials can be represented as

$$(4) \qquad IPM_t = \sum_i \frac{v_{i,1967}}{V_{1967}} \frac{Q_{i,t}}{Q_{i,1967}}$$

where IPM_t is the index of materials production in year t, $v_{i,1967}$ is the value-added by material i in 1967, V_{1967} is the total value-added in 1967, $Q_{i,t}$ is the physical quantity of material i produced in year t, and $Q_{i,1967}$ is the physical quantity of material i produced in 1967.

If one multiplies IPM_t by V_{1967}, this leaves

$$(5) \qquad IPM_t V_{1967} = \sum_i \frac{v_{i,1967}}{Q_{i,1967}} Q_{i,t}.$$

In this form it is clear that to combine ΔN_t and $IPM_t V_{1967}$, one must assume that price per unit of material i in 1967 is proportional to value-added per unit of material i in 1967. That is, for all commodities, one must be able to write

$$(6) \qquad p_{i,1967} = \alpha(v_{i,1967}/Q_{i,1967}).$$

A comparison of the value-added in production and the value of shipments for the various materials indicates that this is not an altogether unreasonable approximation. On the aggregate level, and more roughly on the disaggregate level, price is approximately twice as large as value-added per unit. In the 1971 edition of *Industrial Production*, the Federal Reserve Board calculates both the total value of shipments and the total value-added of the commodities included in their

TABLE 6—EFFECTS OF INCLUDING INVENTORIES, 1954–82

Measure of Volatility	FRB Manufacturing	FRB Materials	Index of Materials Consumed[a]
Mean Cyclical Amplitude of Detrended Series	.1237	.1498	.1389
Standard Deviation of Growth Rates	.0620	.0761	.0690
Standard Deviation of Deviations from Trend	.0610	.0714	.0677

[a] FRB materials index corrected for changes in inventories.

materials classification in 1963. The ratio of the two is 1.97. This ratio can be taken to summarize the relation between price and value-added.

To form the corrected materials index is now straightforward. The materials index is multiplied by 1.97 times the value-added in the production of materials in 1967. The change in real inventories is then subtracted from this figure. This leaves a series on the real value of materials consumed. To return this series to index form, it is normalized by dividing each observation by the value of the series in 1967. Because the inventory data are only available after 1954, an index of materials consumed can only be created for the period 1954–82.

The results of correcting the FRB materials index for movements in inventories are shown in Table 6. The table shows various measures of volatility for both the original materials index and the new materials consumed index. The results are quite strong: the index of materials consumed is substantially smoother than the original FRB materials index. For the mean cyclical amplitude, it is possible to reject the hypothesis that the average amplitudes of the materials consumed and the materials produced indexes are the same at the 95 percent confidence level ($t = 2.66$). Indeed, correcting for inventory movements reduces the discrepancy between the volatility of the materials index and the total FRB manufacturing index by approximately half. This suggests that inventory movements are important and an accurate index of materials consumed is much less volatile than an index of materials produced. For this reason, total output, which is more closely proportional to materials consumed, is less variable than materials produced.

Inventory movements may, in fact, explain more of the difference in volatility between the FRB materials index and the total FRB manufacturing index than the calculations in Table 6 suggest. The crude index of materials consumed presented in Table 6 only corrects for investment in materials inventories. Investment in finished goods inventories may also be important. Goods such as pig iron or lumber are held both as materials inventories by ultimate fabricators and as finished goods inventories by the original pig iron or lumber producers. Since investment in these types of finished goods inventories is also very procyclical, a true index of the materials consumed is probably even less cyclically responsive than one that only takes into account materials inventory investment.

IV. Is the Prewar Index of Industrial Production Excessively Volatile?

Section II shows that both the postwar extensions of Frickey's index are systematically more volatile than the true total index of postwar industrial production. This section extends the analysis to see if the Frickey index is also more volatile than a true index of industrial production would be if it were available for the prewar era. It examines

whether the methods used to form the Frickey index have the same effects in the pre-1914 era as they have in the post-1947 period.

A. *The Relative Importance of Materials in the Prewar and Postwar Eras*

The first issue involved in determining whether the prewar Frickey index is excessively volatile concerns the role of materials in the prewar economy. While it is obvious that the goods included in Frickey's index are for the most part manufactured materials, it is possible that such materials represented a larger fraction of the prewar economy than such materials represent today. If this were true, then Frickey's series might not be as poor an indicator of total industrial production in the prewar era as it is in the postwar era.

To test whether or not this supposition is correct involves devising a measure of how representative a materials index is for both the prewar and postwar economies. One simple measure that can be calculated for both periods is the ratio of the cost of materials to the total value-added in manufacture, where both values are in nominal terms. This ratio provides a rough estimate of the importance of materials in the prewar and postwar eras.

Data on the necessary quantities are available from the *Census of Manufactures* for 1904 and 1967. For both time periods, the cost of materials includes both raw and partially manufactured materials. The resulting ratios of the cost of materials to total value-added in manufacture are 1.26 for 1904 and 1.14 for 1967.[11] The similarity in the ratios suggests that materials are only a slightly larger fraction of the economy in 1904 than in 1967. This implies that the prewar Frickey index and the postwar FRB materials index are approximately equally representative of the underlying economies.

There are, however, some severe limitations to the calculations. While the basic definitions and methods appear to be comparable over time, the data from the two censuses are probably not consistent. Most importantly, because the cost of materials number includes both raw and manufactured materials, this figure involves a substantial amount of double counting. If the degree of vertical integration has changed over time, then there could be different degrees of double counting in the two benchmark years which could affect the calculation. Similarly, because the available data are in nominal terms, relative price changes over time could have affected the ratios.

Nevertheless, this calculation is instructive. It suggests that the importance of materials has not decreased significantly over time. Part of the explanation of this finding is that our definition of materials has changed over time. The manufactured materials classification has come to include increasingly fabricated goods as the economy has become more sophisticated. However, the FRB materials index also includes goods much further along in the production process than does the prewar Frickey index. Thus, the comparison of the cost of materials to total value-added, where the definition of materials changes over time, does provide a legitimate way to assess how representative the prewar and postwar materials indexes are of total industrial production.

B. *The Behavior of Materials Inventories in the Prewar Era*

While the previous comparisons suggest that a materials index is no more representative of total industrial production in the pre-1914 era than in the postwar era, this is still not proof that the prewar Frickey index is excessively volatile. A second issue concerns the behavior of prewar materials inventories. It is possible that materials inventory investment was not procyclical in the prewar era as it is in the postwar era. If this were true, then Frickey's prewar index of materials produced

[11] The data for the calculations for 1904 are from Part 1 of the *Census of Manufactures* (1905, Table 57, p. 109). The data for 1967 are from Vol. 1 of the *Census of Manufactures* (1967, Table 5, p. 45). The results are essentially the same using gross value in the denominator: the ratio of the cost of materials to gross value is .57 for 1904 and .54 for 1967.

might be an adequate proxy for materials consumed and hence for total industrial production.

Historical evidence on the behavior of materials inventories is, unfortunately, very limited. Data on manufactured materials inventories are essentially nonexistent for the pre-1914 period. However, there are a few fragments of inventory data from the interwar period that can be used to test whether the cyclical behavior of inventories is similar in the interwar and postwar eras. It is likely that data from this later, but contiguous, period may reflect prewar inventory movements fairly well. It is difficult to imagine what structural shift could have caused inventory behavior to change dramatically between the early 1900's and the 1920's.

One such fragment comes from Moses Abramovitz's early study of inventories. His study includes data on raw materials inventories for the period 1918–38 (1950, chs. 9 and 10). On the basis of these data, Abramovitz finds that investment in raw materials inventories is decidedly procyclical. Using monthly data and the then fashionable business cycle techniques, Abramovitz concludes that "investment in raw materials inventories tends to conform to cycles in the rate of change in manufacturing activity" (p. 397). While Abramovitz's findings only apply to raw materials inventories rather than to manufactured materials inventories, they are nevertheless suggestive. It seems plausible that the same considerations that govern a firm's choice of raw materials inventories also apply to manufactured materials inventories. Hence, if raw materials inventory investment is procyclical in the prewar era, it is likely that manufactured materials inventory investment is procyclical as well.

A second fragment of data on interwar inventories comes from early issues of the *Survey of Current Business*. Available data on inventories of newsprint at publishers from 1920 to 1937 provide a more direct look at the behavior of an example of manufactured materials inventories. The estimated coefficient of the regression of the change in inventories of newsprint at publishers on the deviations of industrial production from trend is .336 ($s.e = .339$).[12] Because of the very limited sample the standard error is large, but the results are still suggestive of a procyclical relationship.

Finally, a third fragment of evidence concerns the behavior of materials held as finished goods inventories. As mentioned earlier, if investment in materials held as finished goods inventories is procyclical, then cyclical movements in materials produced will be larger than cyclical movements in materials consumed. There is a substantial amount of data on such finished goods inventories of materials for the interwar period. Again, from early issues of the *Survey of Current Business* it is possible to put together a sample of real finished goods inventories from the 1920's and 1930's. The sample includes a variety of materials and intermediate goods, among them lumber, wheat flour, crude petroleum, and steel sheets.[13]

The percentage changes in these series can be pooled and the composite series used to test whether investment in finished goods inventories of materials is procyclical in the interwar period. When the pooled changes in finished goods inventories are regressed on pooled deviations of industrial production from trend, the coefficient is .40 ($s.e. = .15$). This suggests that movements in finished goods inventories are procyclical. It is important to note, however, that the actual correlation between inventory movements and the cycle in this regression is small ($R^2 = .06$). This fact, however, is probably due to the fact that the inventory data are very disaggregated so that industry specific shocks may be dominating the overall influence of the cycle.

[12] The industrial production series used is the conventional FRB index of manufacturing production for 1919–47.

[13] The sample includes 118 observations. The inventory series included in the sample are: Wheat flour (1921–37); Crude petroleum (1923–37); Southern pine lumber (1920–29); Cotton cloth (1926–31); Cotton yarn (1928–31); Refined sugar (1923–37); Steel sheets (1919–35); Paper (1920–31); Woolen cloth (1934–47); Newsprint at mills (1920–37); and Petroleum coke (1923–37).

Taken together, these findings suggest that investment in materials inventories and finished goods inventories of materials are procyclical in the prewar era as they are in the postwar era. As a result, it is likely that Frickey's prewar index of industrial production overstates the volatility of the prewar economy in the same way that the replications of his index overstate the volatility of the postwar economy. Because his methods do not take into account the procyclical movements in the stock of materials, Frickey's original index exaggerates cyclical movements in total manufacturing output in the prewar era.

C. *Volatility of the Shaw-Kuznets Series*

A final, important piece of evidence on the excess volatility of the prewar Frickey index is the behavior of what appears to be a better measure of prewar industrial production. While many prewar indexes of industrial production suffer from the same excess volatility evident in Frickey's series, an annual series created by William H. Shaw (1947) appears to represent cyclical movements more accurately. The Shaw series covers the period 1889–1919. It has been extended back to 1869 by Simon Kuznets (1961).[14]

The Shaw-Kuznets series differs from other historical indexes of industrial production in that it is based on the value of commodity output rather than on the physical quantity of goods produced. To form a real measure of commodity output, the Shaw-Kuznets estimates are deflated by a series of price indexes derived from the Bureau of Labor Statistics Wholesale Price Index. The resulting estimates of real commodity output also differ conceptually from a total index of industrial production in that they include only finished goods.

Because these estimates are not based on quantity data, Shaw and Kuznets have a much larger sample of data with which to work. For a variety of reasons there are far more records on the value of goods produced in the prewar era than on the physical quantity of goods produced. Shaw and Kuznets are able to amass an impressive array of annual data on the value of finished goods produced. Most of these data are from annual state reports and special industry and government publications. These annual state data are used to interpolate between the more comprehensive data on the value of commodity output available in years in which the *Census of Manufactures* was conducted.

The fact that the Shaw-Kuznets series contains a great deal of annual data on the value of finished goods produced suggests that it should be a more accurate measure of prewar industrial production than is the Frickey index which contains only data on materials produced. Because the Shaw-Kuznets series includes data on the value of finished goods such as machines and clothing, it should be free of the excess volatility that comes from using the production of pig iron and cotton to proxy for the output of these finished goods.[15]

Given this reason for believing that the commodity output data represent cycles accurately, it is instructive to examine the volatility characteristics of the prewar Shaw-Kuznets series. It is also useful to consider the volatility characteristics of a postwar continuation of the commodity output series. Fortunately, there exists a postwar series that appears to be very similar to the Shaw-Kuznets series, both conceptually and in its actual calculation. This series is the Federal Reserve Board series on the gross value of finished goods.[16] Like the prewar series, the

[14] For a concise and consistent presentation of the Shaw-Kuznets series, see Kuznets (Table R-21, pp. 553–54).

[15] While the commodity series is free of the excess volatility due to a reliance on materials, Kendrick believes that it may still be systematically flawed because of the deflating process. He states that "the price indexes are usually based on quoted prices and do not take full account of changes in subsidiary terms of sale. 'Net realized' prices tend to fluctuate more than quoted prices over the business cycle and thus the real-product estimates have a downward bias in depressed periods and an upward bias in recoveries" (p. 41).

[16] This series is described in detail in the FRB publication *Industrial Production* (1971, pp. 9–11, and 1976, pp. 29–30).

TABLE 7—VOLATILITY OF COMMODITY OUTPUT

Measure	Shaw-Kuznets (1869–1914)	FRB Gross Value (1954–82)	FRB Final Products (1947–82)
Mean Cyclical Amplitude of Detrended Series[a]	.0937	.0894	.1063
Standard Deviation of Growth Rates	.0567	.0469	.0497
Standard Deviations of Deviations from Trend	.0678	.0345	.0527

[a]Trend commodity output is calculated as the fitted values of a regression of the log of industrial production on a constant and a quadratic trend.

postwar FRB series is a gross value rather than a value-added index. Furthermore, like the Shaw-Kuznets series, this series measures the constant dollar value of only final finished goods.

While the FRB gross value series is a good postwar continuation of the Shaw-Kuznets series, it suffers from two limitations. The most serious of these is that the gross value series is only available after 1954. Since this means that the very volatile late 1940's and early 1950's are excluded, this series will underrepresent the true volatility of commodity output in the postwar era. A second limitation is that the gross value series changes base years in 1967. As a result, in calculating trend commodity output one must estimate the trend over a very short period. As a result the standard deviation of deviations from trend is likely to be artificially low. To deal with these two limitations, I also examine the more common FRB final products index which is available without break from 1947 to 1982. While this is a value-added index, a comparison of this series with the gross value series over the period where they both exist shows that the cyclical properties of the two series are essentially identical.

The various measures of volatility for the prewar and postwar commodity output data are given in Table 7. From these statistics, two characteristics of the commodity output series are obvious. The first is that commodity output has stabilized very little between the pre-1914 and the post-1947 periods. The second is that the Shaw-Kuznets series for 1869–1914 is substantially less volatile than the Frickey index for the same period. The Shaw-Kuznets series is, on average, approximately 35 percent less volatile than the Frickey index. On the other hand, for the postwar era, the two commodity output series are not noticeably less volatile than the FRB index of industrial production in manufacturing.

The fact that the prewar Shaw-Kuznets series is substantially less volatile than the Frickey index provides evidence that the prewar Frickey index is excessively volatile. Since there is reason to believe that the Shaw-Kuznets series is a more accurate measure of the cyclical behavior of prewar industrial production than is the Frickey index, the difference in volatility between the two series suggests that the Frickey series is incorrect.

The fact that the Shaw-Kuznets series shows no stabilization over time also provides support for the view that consistent industrial production data do not show a damping of business cycles between the pre-1914 and the post-1947 eras. When Frickey's index is compared with either its exact postwar replication or the FRB materials index, there appears to be only a slight stabilization. When Shaw's series is compared to a similar postwar series, there appears to be little stabilization of the production of finished goods. In both cases, the dramatic stabilization of the postwar economy apparent in inconsistent data disappears.

V. Conclusions

The preceding analysis yields several conclusions about the historical industrial production series. Section I discussed the possible inconsistencies between the prewar Frickey index and the postwar FRB index of industrial production and described the construction of a postwar index that is consistent with the prewar index. Section II showed that when consistent industrial production data are compared, there is very little damping of business cycle fluctuations between the pre-1914 and the post-1947 period.

The last two sections sought to explain the source of this result. Section III showed that the postwar replications of Frickey's index are more volatile than the true index of industrial production because the Frickey-like series are based very heavily on the production of materials. Because materials inventories are strongly procyclical, an index of materials produced is much more volatile than an index of total industrial production. Section IV showed that the prewar Frickey index is excessively volatile. Data on the relative importance of materials and the behavior of materials inventories in the prewar and postwar eras suggest that the Frickey index is as bad a proxy for total industrial production in the prewar era as a replication of his index is for total industrial production in the postwar era.

It is useful to compare these conclusions on the industrial production series to those I have drawn elsewhere on the historical unemployment and Gross National Product series. The results of the three studies are essentially identical. All three suggest that the prewar macroeconomic data are excessively volatile.

It may seem puzzling that all three studies show similar results. To some degree, the source of the excess volatility in each series is very different. For the unemployment series, the source of exaggerated cyclical fluctuations is the fact that the relationship between unemployment and output is misspecified. The Lebergott unemployment series for 1890–1930 is derived by assuming that the labor force does not move with the cycle and that employment in some sectors moves one-for-one with output in that sector. (See Stanley Lebergott, 1964.) A variety of evidence suggests that these assumptions are false for the prewar era and using them yields a prewar series that is excessively volatile.

The source of exaggerated cyclical fluctuations in the Kuznets *GNP* series involves the specification of the relationship between the available data on commodity output valued at producer prices and total Gross National Product. Kuznets derives estimates of total *GNP* by assuming that *GNP* by sector moves one-for-one with commodity output. However, this assumption, which is certainly false for the postwar era, appears to also be false for the prewar era. *GNP* includes several components that move much less over the cycle than does commodity output. Therefore, using the assumption that the two move together one-for-one yields a series that accentuates cyclical movements.

While the sources of errors in the unemployment and *GNP* data may seem quite different from that in the industrial production data, the problems in each series are, in fact, quite similar. In all three cases the cyclical exaggeration stems from using a series that is too volatile to proxy for the aggregate series being created. In the derivation of the Frickey index, materials production is used to proxy for total industrial production. In the creation of the unemployment series, output is used to proxy for employment. In the construction of the prewar *GNP* data, commodity output is used to proxy for total *GNP*. In all three cases, the aggregate series is assumed to move one-for-one with the series being used as a proxy, rather than substantially less than one-for-one as is almost certainly true. Because all three historical series have essentially the same mistake, it is not surprising that all three are excessively volatile.

While the errors in each series are similar, it is important to note that they are nevertheless independent. That is, the errors identified in one of the series do not cause further errors in the other two series. The reason for this is that none of the three series considered are actually used in the construction of one of the other series. For example, the

output series used to derive the unemployment series is neither the Frickey nor the Kuznets series. Rather, Shaw's series on commodity output is used to construct the unemployment series. Similarly, Kuznets's *GNP* series is based on the Shaw data rather than the Frickey data. Because of this independence, one can identify errors in each series separately.

The fact that the prewar industrial production, unemployment, and *GNP* data are all excessively volatile casts serious doubt on the usual belief that the prewar economy was substantially more volatile than the postwar economy. Indeed, it does appear that the relative stabilization of the postwar economy is a figment of the data. It is important to be very precise about the extent of this conclusion. All three of these studies only examine the data before the Great Depression. As a result, this work in no way challenges the severity of the economic decline of the 1930's. Rather, what this study and its two companions suggest is that the severity of economic fluctuations on both sides of the Great Depression are roughly equal. In fact, one implication of this work may be to emphasize the degree to which the Great Depression is an anomaly in the history of American business cycles.

The conclusion that cycles before and after the Great Depression are equally severe is itself very important. As mentioned in the introduction, this finding contradicts the empirical studies that find a dramatic stabilization between the prewar and postwar economies. On the other hand, this finding may confirm some of the more theoretical analyses of prewar and postwar business cycles. For example, a variety of studies of the prewar economy stress the presence of relatively flexible wages and prices. (See, for example, Jeffrey Sachs, 1980, and Phillip Cagan, 1975.) It is certainly possible that such flexible prices could have enabled the prewar economy to adjust rapidly to various shocks.

The decline in wage and price flexibility between the prewar and postwar eras may explain why the use of both discretionary and automatic stabilizers in the postwar has not yielded a dramatically more stable economy. These stabilizing forces may have served primarily to counteract the possibly destabilizing effects of wage and price rigidity. While this is clearly only a conjecture in need of careful testing, it does suggest an interpretation of the results in this paper that are in accord with relatively Keynesian models of the business cycle.

Thus, the finding that pre-World War I and post-World War II cycles are of equal magnitude need not imply that stabilization policy is ineffective. However, without clear evidence of stabilization, we can no longer simply assert that government stabilization policy is obviously effective. If we wish to argue that policy does matter, then we must posit and test possible explanations for the similarity of the prewar and postwar business cycle.

REFERENCES

Abramovitz, Moses, *Inventories and Business Cycles, with Special Reference to Manufacturers' Inventories*, NBER *Studies in Business Cycles*, No. 4, University Microfilms, 1950.

Burns, Arthur F, "Progress Towards Economic Stability," *American Economic Review*, March 1960, *50*, 1–19.

Cagan, Phillip, "Changes in the Recession Behavior of Wholesale Prices in the 1920s and Post World War II," *Explorations in Economic Research*, Winter 1975, *2*, 54–103.

DeLong, J. Bradford and Summers, Lawrence, "The Changing Cyclical Variability of Economic Activity in the United States," NBER Working Paper No. 1450, September 1984.

Frickey, Edwin, *Production in the United States, 1860–1914*, Harvard Economic Studies, Cambridge: Harvard University Press, 1947.

Kendrick, John W., *Productivity Trends in the United States*, NBER General Studies, No. 71, New York: Arno Press, 1961.

Kuznets, Simon S., *Capital in the American Economy: Its Formation and Financing*, NBER *Studies in Capital Formation and Financing*, No. 9, New York: Arno Press, 1961.

Lebergott, Stanley, *Manpower in Economic*

Growth: *The American Record since 1800*, Economic Handbook Series, New York: McGraw-Hill, 1964.

Lucas, Robert E., Jr., "Understanding Business Cycles," in Karl Brunner and Alan Meltzer, eds., *Stabilization of the Domestic and International Economy*, Vol. 5, Carnegie-Rochester Conference Series on Public Policy, *Journal of Monetary Economics*, Suppl. 1977, 7–29.

Mills, Frederick C., *Economic Tendencies in the United States*: *Aspects of Pre-War and Post-War Changes*, NBER General Series, No. 21., New York: Arno Press, 1932.

Nutter, G. Warren, *The Growth of Industrial Production in the Soviet Union*, NBER General Series, No. 75, University Microfilms, 1962.

Persons, Warren M., *Forecasting Business Cycles*, New York: Wiley & Sons, 1931.

Romer, Christina, "Spurious Volatility in Historical Unemployment Data," *Journal of Political Economy*, February 1986, *94*, 1–37.

______, "The Prewar Business Cycle Reconsidered: New Estimates of Gross National Product, 1872–1918," unpublished manuscript, Princeton University, 1985.

Sachs, Jeffrey, "The Changing Cyclical Behavior of Wages and Prices," *American Economic Review*, March 1980, *70*, 78–90.

Shaw, William Howard, *Value of Commodity Output since 1869*, NBER General Series, No. 48, University Microfilms, 1947.

Stewart, Walter W., "An Index Number of Production," *American Economic Review*, March 1921, *11*, 57–70.

Zarnowitz, Victor and Moore, Geoffrey H., "Major Changes in Cyclical Behavior," NBER Working Paper No. 1395, July 1984.

U.S. Board of Governors of the Federal Reserve System, *Industrial Production*, Washington, 1971 and 1976.

U.S. Bureau of Economic Analysis, *Business Statistics 1977*, Biennial Supplement to *The Survey of Current Business*, Washington, 1978.

______, *Long-Term Economic Growth*, Washington: USGPO, 1973.

U.S. Bureau of the Census, *Historical Statistics of the United States*, Washington: USGPO, 1975.

______, *Census of Manufactures*, Washington: USGPO, 1905; 1967.

The Distributional Welfare Effects of Rising Prices in the United States: The 1970's Experience

By THOMAS M. STOKER*

This paper presents estimates of the distributional welfare impacts of the actual price rises of energy and nonenergy commodities during the 1970–80 decade in the United States. Measures of welfare change based on net compensating variations are computed for different types of families. These show that welfare differences due to changes in commodity prices are minor compared to welfare differences due to different income growth patterns over the decade.

The substantial energy price rises of the 1970's have stimulated a great deal of interest in the reactions and subsequent economic positions of producers and consumers in the U.S. economy. On the consumption side, the major issues concerned how families would react to much higher energy prices, as well as the associated impacts on the welfare of families. Newspaper reports of the time are filled with accounts of how much more difficult life had become for the poor, the elderly, and families on fixed incomes in the face of escalating energy prices, as well as detailed accounts of how much energy budgets had risen for virtually all families.

This paper gives a retrospective view of the U.S. experience, by presenting estimates of the distributional welfare impacts of the actual price rises of the 1970's for different types of families. The first question of interest concerns how widely the welfare impacts vary over families. When energy prices increase, one would expect that welfare impacts will vary over families according to how much energy they consume, which is related to each family's income and demographic makeup. For example, since energy for heating, lighting, and basic transportation is a relative necessity in family budgets, poorer families will tend to experience a greater percentage welfare loss than wealthier families. For a given level of income, larger families will tend to experience a greater percentage loss than smaller families. A greater percentage loss will be experienced by families in less temperate climates, such as the midwestern United States, relative to families in more temperate climates, such as the western United States.

An assessment of welfare differences is necessary for judging the distributional consequences of price changes that are not revealed by an analysis of a "typical" family experience, such as one considering the economic position of a family with per capita income. Moreover, if such welfare differences are substantial, their recognition is necessary for formulating more equitable policy responses to soften the impact of price shocks, whether the price changes arise from outside economic events, such as OPEC-led price increases, or policy changes that affect either domestic energy pricing or applicable energy tariffs.

The second question concerns the appropriateness of focusing policy concerns on the behavior of the price of a single commodity, such as energy. A large amount of the national policy discussion of the 1970's was focused on energy, which may seem justifiable because the nominal price of energy more than tripled during the decade. However, nominal prices for food and clothing, housing services, and consumer services almost

*Associate Professor of Applied Economics, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139. This research was funded by a grant from the MIT Center for Energy Policy Research. I thank D. Jorgenson, R. Pindyck, R. Schmalensee, and D. Wood for helpful comments, and K. McLeod for expert research assistance.

doubled, and nominal prices of other goods in the budget more than doubled. Also relevant is the nominal income history of individual families over the decade—whether income is gauged to the level of prices, fixed in nominal terms, or following the cyclic behavior of average income in the U.S. economy. While plain economic sense dictates that a family's reaction to changes in all prices and income is important for its welfare, this paper numerically points out the differences between welfare measures based solely on energy expenditures and welfare measures that account for all price and income changes.

For measuring welfare impacts with rising energy prices, a natural technique is to take the initial quantity of energy consumed by a family prior to the price shock, revalue it at the increased price, and measure the welfare loss as the additional cost of the initial quantity. While a revaluation technique of this type provides a measure of how much more expensive energy has become, it fails to account for several important features of the economic behavior of families. The substitution behavior of families is ignored, whereby the quantity purchased of energy is reduced via both the own-price effect of reducing energy purchases because of the increased price as well as the cross-price effect of substitution away from energy toward relatively less expensive consumption items. Any change in the level of family income (or total expenditure budget) is also ignored. By not accounting for these features, revaluation measures are likely to significantly misstate the actual losses (or gains) experienced by individual families.

A useful measure of the welfare impacts on families is the net compensating variation,[1] which accommodates both substitution behavior and changes in family budgets. The compensating variation arising from price changes refers to the amount of money that one would have to pay a family under the new prices to make it exactly as well off as it was under the initial prices. This measure can be positive or negative according to whether the family is worse off or better off under the new prices, and will vary with the level of income and demographic makeup of the family. The net compensating variation is the difference between the compensating variation and the change in total expenditures.

In this paper, net compensating variation measures of welfare change are presented corresponding to the observed annual price experience in the United States over the period from 1970 to 1980. Welfare measures are computed for families with different demographic makeups and different total expenditure budget histories over the decade, or, in short, different total expenditure profiles. Five demographic dimensions are included: family size, age of head, region of residence, race of head, and urban or rural residence. Total expenditure profiles include initial (1970) total expenditures ranging from \$5,000 to \$25,000, with several growth patterns over the decade: growth constant in real terms (i.e., proportional to a Consumer Price Index), proportional to per capita income, and at nominal rates ranging from 0 percent (fixed nominal total expenditures) to 20 percent.[2]

The overall conclusions of the results are as follows. The revaluation measures of energy purchases as well as compensation measures based solely on energy prices bear no relation to the correct welfare change measures. While substantial demographic differences exist in demand behavior, the demographic differences in welfare impacts are slight, with a family whose total expenditures are constant in real terms realizing gains of 6–7 percent of the initial budget, due to substitution in 1980 relative to 1970. Large losses in welfare position occur only for

[1] Compensating variation measures of welfare gain or loss were introduced by John Hicks (1939, 1942).

[2] In this study I make no attempt to combine individual compensation measures into an overall social welfare or inequality index. Such combinations of compensation measures are discussed in Robert Willig (1981) and Willig and Elizabeth Bailey (1981). John Muellbauer (1974a, b and 1978) and Dale Jorgenson and Daniel Slesnick (1984) present analyses of welfare inequality using social welfare functions that include the influence of differing household characteristics.

families whose nominal expenditure budgets grew at rates substantially below the inflation rate. Consequently, this study reaffirms the standard economic position that problems of families under economic duress, whether caused by rising energy prices or other prices, should be treated as problems of insufficient income and dealt with, without special attention to any particular price.

I. Methodology

A. *Family Demand Behavior and Net Compensating Variation*

In order to compute measures of changes in family welfare that account for substitution among commodities, a complete empirical description of family demand behavior is required. Such a demand model must characterize the effects of changes in all commodity prices, income and the demographic makeup of individual families and also be consistent with utility maximization by each family, so that tradeoffs between different commodity consumption levels are consistently characterized. In this section I introduce a model of U.S. family demand behavior that fulfils these criteria (the translog model of Dale Jorgenson, Lawrence Lau, and myself, 1982; hereafter J-L-S) and present the associated formulae to be used for calculating compensating variation welfare measures.[3]

The J-L-S demand model is based on explicit modeling of the connection between individual family demand functions and economywide aggregate commodity expenditures. The system is estimated by pooling cross-section data on expenditures by individual families for a single year, together with aggregate annual time-series data on expenditures, price levels, and statistics of the joint distribution of family income and demographic variables. To my knowledge, the J-L-S system represents the only available empirical model of U.S. family demand behavior that fulfils the criteria required for this study.[4]

The J-L-S demand model describes family budget allocation to five overall categories of commodity expenditures, indexed by $n = 1,\dots,5$, given as[5] 1) Energy; 2) Food and Clothing; 3) Other Nondurable Goods; 4) Capital Services; and 5) Consumer Services.

The annual time periods are indexed by t, with $t = 0,\dots,T$, and families are indexed by k, with $k = 1,\dots,K_t$, where K_t is the population size in year t. In addition to total budget expenditures (which represents income for my purposes), families are differentiated along the following demographic dimensions:

1) Family Size: 1, 2, 3, 4, 5, 6, 7 or more persons;

2) Age of Head: 15–24, 25–34, 35–44, 45–54, 55–64, 65 and over;

3) Region of Residence: Northeast, North Central, South, and West;

4) Race of Head: White, Nonwhite;

5) Type of Residence: Urban, Rural.

These dimensions are represented by qualitative variables (so as to not constrain the effects of changes along each dimension), that allow for 672 different types of families in the population at any given total expenditure level.

[3] The technique of computing compensating variation welfare measures from the J-L-S translog demand model was first applied in our 1980 article.

[4] Part of the reason for this is that the approach of combining individual cross-section and aggregate time-series data sources provides the only practical solution for simultaneously estimating individual effects of prices, income and a realistic number of demographic influences given the available data for the United States. For more general modeling of all of the important effects, data on expenditure patterns of individual families that exhibit independent variation of prices, income and all demographic dimensions of interest are required, such as would be observed in a multiyear panel study of family budgets. Unfortunately, data of this type containing full coverage of family budget categories are not presently available for families in the United States.

[5] See our study (1982) for discussion of the data categories. Capital services refers to the service flow from consumer durables and housing, whereas investment in these categories is treated as altering the stock of capital (as well as replacement). These service flow data were constructed from data on the stocks of housing and consumer durables via the perpetual inventory method described in Christensen and Jorgenson (1970) and Jorgenson (1986).

The notation used for the formulae is as follows. The price of the nth commodity category in year t is denoted as p_{nt}, with the full vector of prices denoted as $p_t = (p_{1t},\dots,p_{5t})$. The quantity purchased of commodity n by family k in year t is denoted as x_{nkt}, with budget share $w_{nkt} = p_{nt}x_{nkt}/M_{kt}$, where $M_{kt} = \Sigma_n(p_{nt}x_{nkt})$ is total expenditure by family k in year t. The values of the qualitative demographic variables for family k in year t are denoted as A_{skt}, where s indexes demographic types, and $s=1,\dots,16$ for the J-L-S model. The symbol A_{kt} represents the vector of all qualitative demographic variables.

The equations of the J-L-S demand model posit that the nth budget share for family k in year t takes the form:

$$(1)\quad w_{nkt} = \frac{1}{D(p_t)}\left[\alpha_n + \sum_j \beta_{nj}\ln p_{jt} - \beta_{Mn}\ln M_{kt} + \sum_s \beta_{Ans}A_{skt}\right] \qquad n=1,\dots,5,$$

where α_n, β_{nj}, β_{Mn}, and β_{Ans} are preference parameters that are estimated for n, $j=1,\dots,5$, $s=1,\dots,16$, and $D(p_t) = -1 + \Sigma\beta_{Mn}\ln p_{nt}$.[6] For purposes of overall interpretation, the system of equations (1) describes budget shares as a nonlinear function of prices, and for any given level of prices, budget shares are a linear function of log-total expenditure and the qualitative demographic variables. Nonlinearity of Engel curves (i.e., nonhomotheticity of preferences) is represented by budget shares that vary with total expenditure size, which is implied by nonzero values of β_{Mn}, $n=1,\dots,5$. Demographic differences in demand behavior are represented as level differences in budget shares, implied by nonzero values of β_{Ans}, $n=1,\dots,5$, $s=1,\dots,16$.

For the detailed parameter estimate results and associated analysis, the reader is referred to the original J-L-S paper (1982).[7] Overall features of the results imply substantial nonlinearities in the Engel curves for all commodities, with the capital share increasing with total expenditure, and all other shares decreasing. Demographic effects are precisely estimated, and in many cases represent large effects on budget share allocations. These two features underlie many of the differences in the welfare measures to be presented.

For an ordinal representation of the preference levels of each family, an indirect utility function which is consistent with the share equations (1) is given as

$$(2)\quad \ln V_{kt} = \sum_n \ln p_{nt}\left(\alpha_n + (1/2)\sum_j \beta_{nj}\ln p_{jt} + \sum_s \beta_{Ans}A_{skt}\right) - D(p_t)\ln M_{kt},$$

where V_{kt} represents the utility level of family k at year t, which is a quadratic function of log-prices, log-total expenditure, and demographic variables. The differing economic needs associated with demographic differences are represented as fixed effects on the price elasticities of indirect utility V_{kt}.[8] The minimum expenditure function implied by the indirect utility representation (2) is given as

$$(3)\quad \ln M(p_t, V_{kt}, A_{kt}) = \frac{1}{D(p_t)}\left[\sum_n \ln p_{nt}\left(\alpha_n + (1/2) \times \sum_j \beta_{nj}\ln p_{jt} + \sum_s \beta_{Ans}A_{skt}\right) - \ln V_{kt}\right].$$

[6]For estimation, the parameters of equations (1) are subject to integrability restrictions, which assure that (1) is consistent with preference maximization for each family. These restrictions include adding-up and Slutsky symmetry constraints as well as inequality constraints which assure the negative semidefiniteness of the matrix of compensated price effects (see J-L-S, 1982, for details). The resulting estimating equations contain 82 parameters to be estimated.

[7]Table 3; pp. 212–15 contains the parameter values used.

[8]Equation (2) is ordinal to the extent that the actual preferences of each family can differ from (2) by a family-specific monotonic transformation. Because of this, values of utility V are not comparable across

The Hicksian notion of compensating variation is used here for making welfare comparisons across families, and is defined as follows. In the base period $t=0$, welfare of the family k is determined by initial prices (p_0), total expenditure (M_{k0}), and demographic makeup (A_{k0}). In a subsequent period $t=1$ (say), prices change to p_1. The compensating variation associated with the price change is defined as

$$(4)\quad CV_k(p_0,p_1)=M(p_1,V_{k0},A_{k0})-M_{k0},$$

or the amount of money (positive or negative) that one would have to give the family in order to make them just as well of as in the initial period. Since total expenditure may also change from period 0 to period 1, define the net compensating variation as the difference between the compensating variation and the change in total expenditure as

$$(5)\quad NCV_k(p_0,p_1) = CV_k(p_0,p_1)-(M_{k1}-M_{k0}),$$

where M_{k1} is the period $t=1$ total expenditure value. If the net compensating variation is negative, welfare of the family is increased by the price and total expenditure change; if positive, welfare is decreased. Dollar value comparisons of welfare changes across families can proceed on the basis that a larger net compensating variation value indicates a family to which a greater amount of money must be given to maintain the initial welfare level, and in that sense, such a family is "worse off" than the family with the smaller value.

While theoretically correct, computed values of the net compensating variation (5) can be difficult to interpret because negative values correspond with welfare gains and positive values correspond with welfare losses. Therefore, for ease in interpretation, the tables of Section II report dollar values of welfare change, defined as the negative of the net compensating variation

$$(6)\quad C_k=-NCV_k(p_0,p_1).$$

The sign of C_k corresponds with the direction of welfare change—positive values indicate gains and negative values indicate losses. Some further conceptual basis for using C_k exists: namely when the total expenditure budget (and configuration of demographic attributes) is unchanged between periods 0 and 1, C_k corresponds with the change in "money metric utility" defined with p_1 as reference price vector.[9]

Compensating variation measures of welfare change are based on comparisons made with preferences (including demographic makeup) held constant at base period levels. An alternative measure can be defined for varying prices with preferences held constant at levels prevailing after the price changes; the Hicksian notion of equivalent variation.[10] Since the application at hand is to measure the welfare position of families after the actual price changes experienced during the 1970's, the above compensating variation

families without further assumptions, and no such comparability is used in this study. Such comparability is required to measure welfare differences along demographic dimensions, such as how much expenditure difference makes a small family as well off as a large family —see Robert Pollak and Terence Wales (1979) on this point.

[9]This is pointed out in Angus Deaton and Muellbauer (1980, p. 186) as follows. The money metric utility with reference price p_R is defined as $M(p_R,V_k,A_k)$, which is an increasing function (i.e., index) of V_k. When total expenditure is the same in period 1 and period 0, then $M(p_1,V_{k1},A_k)=M(p_0,V_{k0},A_k)$. Thus for reference price vector $p_R=p_1$, the change in money metric utility is

$$\begin{aligned}M(p_1,V_{k1},A_k)-M(p_1,V_{k0},A_k) &= M(p_0,V_{k0},A_k)-M(p_1,V_{k0},A_k)\\ &= -CV_k=C_k.\end{aligned}$$

[10]Net equivalent variations correspond with changes in money metric utility with reference price p_0, and thus may be preferable for certain applications, such as comparing the welfare impacts of different policies yielding several alternative price paths (since the reference price p_0 does not vary with alternative values of p_1). See Deaton and Muellbauer for discussion, as well as Paul Samuelson (1974) and John Chipman and James Moore (1976, 1980).

concepts are used for ease of interpretation, with 1970 as the base year for comparison.

Two "single price" welfare cost measures are also presented for comparison, that focus solely on changes in the price of energy. Take energy as commodity 1, with price p_{1t} at time t, $t=0,1$. A typical revaluation measure of the cost of the energy price change from p_{10} to p_{11} is the change in cost of initial energy purchases:

$$R_k = (p_{11} - p_{10}) x_{1k0}. \tag{7}$$

The evaluation R_k fails to represent overall welfare change because it ignores substitution behavior, changes in nonenergy prices and changes in total expenditures.[11] Substitution behavior is accommodated in the other single price measure; the compensating variation associated with the energy price change. Let $p_1^* = (p_{11}, p_{20}, \dots, p_{50})$ be the price vector with the final energy price but initial values of other prices, and define

$$S_k = CV_k(p_0, p_1^*). \tag{8}$$

The measure S_k is a partial cost measure[12] that can be compared to the revaluation R_k. Finally, it should be kept in mind that R_k and S_k are cost measures of price-change impacts—they utilize the sign convention of the compensation measure NCV_k, opposite to that of welfare change C_k.

[11] The revaluation R_k can act as a bound on the "single price" compensation measure (8): see my 1985 paper for discussion of this as well as many classical references to this property.

[12] I do not use traditional (area) consumer's surplus measures because they only represent approximations to the exact compensating variation values, due to the presence of income effects. While the conditions under which consumer's surplus exactly equals the compensating variation are quite strict (see Chipman and Moore, 1976), the approximation can be quite close in nonpathological circumstances (see Willig, 1976). In this regard, Jerry Hausman (1981) has devised a method for constructing preference levels from empirically estimated demand functions, which considerably lessens practical interest in consumer's surplus welfare measures.

B. *Data Utilized*

The data utilized for the calculations are presented in the Appendix. The nominal price data are constructed from the corresponding series for personal consumption expenditures, with the exception of the series for the price of capital services, which was constructed separately to maintain comparability with the data used to estimate the J-L-S model.[13] All "real" expenditure profile scenarios are constructed with respect to the Adjusted Consumer Price Index (ACPI), which is constructed by modifying the standard Consumer Price Index (CPI) to properly include the price of capital services and remove prices of durable goods. I include the corresponding data on the standard CPI for comparison with the ACPI data. All nominal and real price series are normalized to 1 in 1972.

II. Measures of the Welfare Changes of U.S. Families: 1970–80

This section presents measures of the welfare gains and losses experienced by different types of families in the United States due to variations in the levels of prices from 1970 through 1980. Part A gives measures corresponding to different family total expenditure profiles and Part B gives measures corresponding to different family demographic configurations.

A number of conventions are utilized in the presentation of the results, which are summarized as follows. All dollar values are annual measures presented in 1972 (constant) dollars, to facilitate comparisons across years, which are easily convertible to another base year by the use of the appropriate ACPI value. While welfare measures can be computed for any family income and demographic configuration, for comparison I utilize a base family with 1970 total expenditure of $10,000, family size of 3, age of head

[13] The capital services price data are imputed prices on annual flow measures of capital services, described in fn. 5. The 1970–80 capital services price and quantity values were graciously provided by Dale Jorgenson.

35–44, Northeast urban residence, and white race. The impacts of variation along a particular dimension will be analyzed, while holding the other dimensions constant at their above-base values.

A. *The Effects of Varying Total Expenditure Profiles*

This section studies the effects of family income history by presenting welfare measures that correspond to differing total expenditure profiles over 1970–80. Each expenditure profile can be interpreted as an after-tax income profile, provided that savings are a stable (family specific) fraction of after-tax income. Three broad issues are addressed by the results of this section. First, how large are the welfare changes experienced by families with "typical" total expenditure profiles? Toward this end, Table 1 presents annual welfare comparisons to 1970 for constant real total expenditures and expenditure growth patterns that follow per capita income. Second, how do the welfare measures vary for families whose total expenditures grew slowly relative to those whose total expenditures grew quickly? Table 2 gives 1970–80 measures corresponding with constant nominal total expenditure growth rates from 0 to 20 percent. Third, what is the effect of differing initial (1970) total expenditure levels on the welfare measures? Table 3 gives 1970–80 measures corresponding to different initial expenditure levels, each constant in real terms over the 1970–80 period.

Table 1 presents the welfare effects of the base family with "typical" total expenditure growth profiles. Column 4 gives the revaluation measure of the additional cost of 1970 energy purchases. This series clearly reflects the enormous jumps in real energy prices of the 1973–74 and 1978–80 periods, as well as the gradual real declines in energy prices over the 1970–72 and 1976–78 periods. These features are also reflected in column 5, the (partial) compensating variations of the energy price changes.

Column 1 of Table 1 gives the welfare change measures for total expenditures, held constant in real terms over the period—nominal total expenditures change proportionately with the ACPI. Note that relative to 1970, the typical family was worse off in 1972–74 and better off from 1975 to 1980, although the gains in 1977 and 1978 are negligible. The most substantial improvement in position relative to 1970 occurs from 1974 to 1975 and 1978 to 1980.

In order to understand the source of these variations, recall first that total expenditures are being held constant in proportion to the ACPI. Initially, one might jump to the conclusion that all of the welfare change values should be positive because the ACPI is basically a fixed weight price index (with weights revised every 10 years); however, this conclusion is unwarranted. The standard argument that fixed weight indexation will overcompensate (because the family can substitute, readjusting its budget shares optimally in the face of relative price shifts) requires that the weights of the price index correspond with the initial optimal budget shares of the family. Because the welfare change measures are computed for a particular family, and the weights in the CPI and the ACPI reflect average shares in the economy, there is no reason why the 1970 optimal budget shares for our typical family must coincide with the average shares. On this point, the J-L-S model implies 1970 budget shares for the typical family for energy, food and clothing, capital services, consumer services, and other nondurables of 3.7, 21.1, 34.3, 25.5, and 15.2 percent, respectively, whereas the 1970 aggregate budget shares of these commodity groups are 4.6, 21.5, 26.9, 27.5, and 19.4 percent, respectively.

With this in mind, the source of the various gains and losses becomes apparent. Namely, the losses in welfare in 1972 and 1973 occur because of the rise in real capital prices, which in spite of substitution behavior, is not sufficiently compensated for by expenditure growth proportional to the ACPI. The changes in position in 1974–75 reflect the substitution away from energy and other nondurables and into capital services and consumer services. Finally, the gains in welfare for 1979–80 are similarly due to substitution away from energy and other nondurables into food and clothing, capital

TABLE 1—ANNUAL WELFARE CHANGES FROM 1970 TO 1980
(1972 Dollars)

	Welfare Change			Partial Cost Measures	
Year	Real Expenditure Constant (1)	Per Family Expenditure Growth (2)	Per Capita Expenditure Growth (3)	Extra Cost of 1970 Energy Quantities (4)	Compensating Variation of Energy Price Increase (5)
1970	0	0	0	0	0
1971	13.08	145.08	325.25	−3.60	−3.66
1972	−73.7	460.47	970.73	−16.79	−17.34
1973	−31.9	559.63	1428.07	−15.47	−16.19
1974	−17.24	268.92	1025.04	55.15	48.42
1975	254.14	335.31	1200.10	65.75	56.57
1976	79.22	494.97	1424.67	62.10	53.75
1977	0	587.91	1743.27	59.51	51.73
1978	−0.42	712.12	2198.01	50.60	44.79
1979	150.17	474.01	2308.05	104.64	83.38
1980	648.31	510.57	2234.72	177.66	124.33

services, and consumer services. This corresponds with the strong increases in real prices for energy and other nondurables relative to the decreases in real prices for food and clothing, consumer services, and particularly capital services.

These figures are appropriate for considering families whose total expenditures are gauged to standard cost of living adjustments, but this may not represent a sensible "typical family" scenario. Toward this end, columns 2 and 3 of Table 1 display measures that correspond to total expenditure growth that follows U.S. average income paths.

Column 2 gives the annual welfare change measures for the base family whose total expenditures vary proportionally with average income per family. Each year again reflects an improved position relative to 1970, although relatively smaller improvements are registered for the recession years of 1974–75 and 1979–80. Substantial substitution gains are again evidenced for 1980, as per family real average income falls below the 1970 level, but the net welfare position is improved.

While per family average income may on the surface appear to be a good measure of typical experience, its definition renders this interpretation somewhat misleading. Because unrelated individuals (families of size 1) are counted with the same weight as other families, a fall in average family income can be due to a fall in average family size. In particular, from 1970 to 1980, the share of unrelated individuals in total households increased from 23 to 31 percent, whereas the share of unrelated individuals in the total (individual) population increased only from 8 to 12 percent. Consequently, the fall in real average family income may just be an artifact of the changing demographic mix of the United States, and not related to the income positions of individual families. To correct

TABLE 2—WELFARE CHANGES FROM 1970 TO 1980 FOR DIFFERENT NOMINAL RATES OF EXPENDITURE GROWTH (in 1972 dollars)

Nominal Percentage Growth Rate	Welfare Change from 1970
0	−4745.33
2.5	−3455.16
5	−1818.42
6	−1102.40
7	− 290.29
7.5	142.15
8	593.09
9	1553.22
10	2596.01
12.5	5606.62
15	9283.58
20	19169.60

TABLE 3—WELFARE CHANGES FROM 1970 TO 1980 FOR DIFFERENT REAL EXPENDITURE LEVELS (in 1972 dollars)

			Partial Cost Measures	
Real Total Expenditure Level	Welfare Change from 1970	Percent of Total Expenditure Level	Extra Cost of 1970 Energy Quantities	Compensating Variation of Energy Price Increase
5,000	264.19	5.28	117.93	86.17
8,000	487.89	6.09	157.10	111.81
10,000	648.32	6.48	177.63	124.33
12,000	815.53	6.79	194.78	134.08
15,000	1076.65	7.17	215.36	144.51
20,000	1533.47	7.66	238.81	153.06
25,000	2011.35	8.04	251.64	152.96

for this, column 3 presents measures computed on the basis of total expenditures adjusted proportionately to real per capita income, where averaging is performed over the total population of individuals, not households. This scenario corresponds to total expenditure growth gauged to the number of individuals in the family as well as economywide income growth. Here the welfare measures indicate much larger improvements, corresponding to the net effects of substitution together with the substantial real growth in per capita income.

A depiction of the welfare impacts on families with high versus low total expenditure profiles is presented in Table 2, where 1980 welfare change measures are presented for incomes growing at constant nominal rates, so, for example, 0 percent corresponds to fixed nominal income over 1970–80. Here, low levels of nominal income growth register substantial losses, which swamp the gains available from substitution. The break-even nominal growth rate is 7.33 percent, which is substantially smaller than the growth rate of 8.06 percent that would result in the same real (ACPI) total expenditure in 1980 as in 1970. It should be kept in mind that the majority of this difference is due to the substitution gains made possible by the relative price shifts of the 1978–80 period. Large growth rates correspond with large gains, again swamping the substitution effects.

Table 3 compares the welfare changes experienced by the base family with varying 1970 initial total expenditures, where each total expenditure level grows proportionately with the ACPI. Because ACPI total expenditure adjustment is utilized, all measures register gains, with larger gains accruing to families with larger initial total expenditure budgets. Since energy is not an inferior good, the expense of 1970 quantities also rises with total expenditure. Moreover, since energy is not a luxury good, one might expect that substitution gains would decline as a percentage of total expenditure. However, the opposite occurs, with the welfare change measures representing 5.3 percent of total expenditure for families with total expenditures of $5,000, rising uniformly to 8.0 percent for families with total expenditure levels of $25,000. The reason for this is the substantially larger purchases of capital services by richer families, with the gains from exploiting the decreased price of capital services dominating the benefits from shifting away from energy.[14]

The overall implication of these tables is that welfare measures only indicate large losses for families whose total expenditures

[14] Note that the partial cost measures do reflect welfare changes that decline in percentage terms. In particular, the compensating variation of the energy price change represents 1.7 percent of total expenditures for families with $5,000 budgets, declining to .6 percent of total expenditures for families with $25,000 budgets.

TABLE 4—WELFARE CHANGES FROM 1970 TO 1980 FOR DIFFERENT FAMILY SIZES (in 1972 dollars)

Family Size	Welfare Change from 1970	Partial Cost Measures: Extra Cost of 1970 Energy Quantities	Partial Cost Measures: Compensating Variation of Energy Price Increase
1	813.40	88.91	52.18
2	697.22	161.39	110.99
3	648.32	177.63	124.33
4	614.07	191.59	135.80
5	605.70	191.97	136.11
6	571.68	206.58	148.14
7 or more	541.59	199.53	142.33

seriously lagged behind inflation, with the substitution gains small by comparison to losses incurred by families having to rely on slowly growing or fixed nominal total expenditures. The measured benefits to substitution vary quite a bit over the 1970–80 period, with the majority of the exploitable substitution gains occuring during the last (1978–80) period of large energy price shocks, combined with real declines in the prices of food and clothing, consumer services, and especially capital services. The substitution gains themselves vary with initial total expenditure position in a predictable way, with low-income families enjoying a smaller percentage benefit from the substitution away from energy toward other commodities, primarily capital services.

B. *Varying Demographic Structure*

In this subsection, measures of welfare change are presented that correspond to families of varying sizes, ages, regions of residence, type of residence, and race. There are two broad issues addressed by these measures. First, since families of differing demographic makeup have different needs and different observed consumption patterns, the welfare changes inherent to the 1970–80 price experience may differ dramatically. Second, if such welfare changes do vary substantially, then the typical family scenarios of Tables 1–3 are not very informative. In this case, an intelligent view of the overall welfare impact picture would require explicit recognition of both the income and demographic positions of families in the economy. As with Tables 2 and 3, for simplicity I focus on 1970 to 1980 welfare comparisons.

The effect of varying family size is presented in Table 4. The welfare change measures decline with family size, indicating smaller welfare gains for larger families. This occurs because larger families purchase larger amounts of food and other commodities, and smaller amounts of capital (housing) services, and therefore benefit less from the lower price of capital. In other words, given the actual price changes, larger families benefit less because of their heavier reliance on food, clothing, and miscellaneous commodities in the budget. This should not be taken to say that smaller families are better off in an absolute sense than larger families, but just that smaller families at a given real total expenditure level experienced more substitution benefits than larger families at the same level of total expenditures. It is natural to expect that a one- or two-person family with a $10,000 total expenditure budget enjoys a higher economic standard of living than a family of seven or more persons with a $10,000 budget, however, no comparability of this type is utilized in the calculations.

The welfare change measures corresponding to families with different ages of head are presented in Table 5. The substitution benefits are the greatest for the youngest and oldest family types, with the lowest benefits accruing to the middle-aged 35–64 group, although the differences between the age

TABLE 5—WELFARE CHANGES FROM 1970 TO 1980 FOR DIFFERENT AGES OF HEAD (in 1972 dollars)

Age of Head	Welfare Change from 1970	Partial Cost Measures: Extra Cost of 1970 Energy Quantities	Partial Cost Measures: Compensating Variation of Energy Price Increase
15–24	711.35	160.56	110.32
25–34	664.55	167.90	116.34
35–44	648.32	177.63	124.33
45–54	631.34	186.87	131.92
55–64	635.23	199.00	141.91
65+	674.62	195.12	138.77

TABLE 6—WELFARE CHANGES FROM 1970 TO 1980 FOR DIFFERENT REGIONS OF RESIDENCE (in 1972 dollars)

Region	Welfare Change from 1970	Partial Cost Measures: Extra Cost of 1970 Energy Quantities	Partial Cost Measures: Compensating Variation of Energy Price Increase
Northeast	648.32	177.63	124.33
North Central	636.71	217.87	157.44
South	609.63	212.27	152.81
West	695.38	157.91	108.02

groups are small. This is again because of a higher reliance by middle-age families on food, clothing, and miscellaneous commodities, and a smaller reliance on housing, with less benefits possible from the decrease in the capital price. As with the discussion of family-size effects, it must be kept in mind that these results focus only on age effects, holding total expenditure constant at $10,000, and therefore do not incorporate any income growth with advancing age.

Measures for differing regions are presented in Table 6. The smallest benefit to substitution occurs for families in the South region of the United States, with the largest benefit for families in the West. The differences are quite small, and appear to be due to smaller expenditure shares to consumer services by families in the South, together with slightly larger shares devoted to other nondurable goods.

Race effects are profiled in Table 7, where marginally greater substitution benefits are shown for nonwhites than whites. The consumption patterns between families with white and nonwhite heads differ only by a slightly higher capital share for whites compared to slightly lower shares for all other nonenergy commodities. The relative race benefit differences for families with the same total expenditure are so slight, however, that the joint effects of decline in prices for food and clothing, consumer services, and capital services relative to energy and other commodities give the same net gains for each type of family.

The urban-rural residence distinction is profiled in Table 8, with rural families showing a much smaller benefit to substitution than urban families. This is surprising because rural families devote a much higher share of total expenditures to capital and

TABLE 7—WELFARE CHANGES FROM 1970 TO 1980 FOR DIFFERENT TYPES OF RACE (in 1972 dollars)

Race	Welfare Change from 1970	Partial Cost Measures: Extra Cost of 1970 Energy Quantities	Partial Cost Measures: Compensating Variation of Energy Price Increase
White	648.32	177.63	124.33
Nonwhite	653.35	160.82	110.52

TABLE 8—WELFARE CHANGES FROM 1970 TO 1980 FOR DIFFERENT TYPES OF RESIDENCE (in 1972 dollars)

Type of Residence	Welfare Change from 1970	Partial Cost Measures: Extra Cost of 1970 Energy Quantities	Partial Cost Measures: Compensating Variation of Energy Price Increase
Urban	648.32	177.63	124.33
Rural	563.68	259.61	191.89

energy, and much lower shares to food and clothing and consumer services than do urban families. Therefore, the net results of the 1970–80 relative price differences suggest that the benefits from the decline in food and clothing and consumer service prices outweighed the benefits from the lower capital price and (substitution away) from the higher energy prices. While it is understandable that rural families would benefit less from real declines in food and clothing prices, because of food and clothing produced in the home, the net results are still surprising as to the lack of benefits from both the capital price decline and substitution away from energy.

In overview, since families of differing demographic profiles display quite different consumption patterns, the welfare experiences of such families differ significantly. But these welfare differences are by and large very small, with only the large family/small family distinction rendering differences in excess of 1 percent of total expenditure ($100). Consequently, these differences appear very minor in comparison to the measures corresponding to the different total expenditure growth profiles of Table 2. Therefore, while the demographic effects on consumption patterns are important, the most substantive welfare changes due to the U.S. price experience from 1970 to 1980 appear much more directly related to differing total expenditure profiles over the decade.

III. Conclusion

This paper presented net compensating variation welfare measures of the changes in the economic position of U.S. families from 1970 to 1980. The computations are based on an empirical description of individual family demand behavior, the translog demand model of J-L-S. The effects of varying income levels, income profiles, family size, age of head, region of residence, race and urban or rural residence are illustrated and discussed.

While every family in the United States has experienced a particular set of changes along some or all of these dimensions during the 1970–80 decade, the welfare change mea-

sures presented in this paper give a fairly clear depiction of the important sources of welfare gain or loss, namely income profile. While families differ significantly in the amounts of energy that they consume, by far and away the greatest losses or gains are associated with overall income variations. Families whose nominal incomes are growing at a 0–5 percent rate are injured dramatically when inflation is proceeding at a 7–8 percent rate, independently of the source of the price changes. While families with certain demographic attributes may benefit relatively more from substitution between commodity categories, such benefits are small relative to the overall losses or gains of purchasing power.[15] It is unlikely that the relative distribution of welfare levels is much affected by the substitution possibilities afforded by the actual price changes from 1970 to 1980.

While the exact values of the welfare gains presented above depend on the precise empirical characterization of family preferences, one can argue that the broad conclusions above apply more generally. In particular, the basic demand modeling approach used here takes as a proviso that expenditure levels can be adjusted costlessly at the margin, which may be unrealistic for commodities such as housing services. This objection suggests that the actual substitution possibilities available to individual families could be less flexible than those implied by the estimated preference parameters. But in this case, the actual welfare gains available from substitution would be smaller than those presented above. Since there is no obvious reason why certain demographic groups would face dramatically larger adjustment costs than other groups, there is likewise no obvious reason to think that estimated differences in welfare impacts across demographic groups are too small. Consequently, the broad conclusions about the relative distribution of welfare levels are not likely affected by the presence of costs of substitution, or any other impediments that render family substitution possibilities less flexible.

In conclusion, the results of this paper computationally underline the standard economic policy prescription in the face of price changes, namely that the effects of all price changes are important. When the price of a particular commodity gains attention because of rapid escalation, as with energy in the 1970–80 decade, it is not sensible to focus any policy decisions to affect the welfare of families on only that price in isolation. All prices must be simultaneously considered, together with the relative importance of the associated commodities in the budget. Transfer mechanisms attached to one price, or policies aimed at altering only a single price, will necessarily ignore the true economic opportunities facing individual families. To illustrate this latter point, suppose that the base family will receive a transfer payment equal to the additional cost of its 1970 energy purchase, which from Table 1 is \$178 (1972 dollars). If the family's total expenditure has remained constant in real terms, then the family is better off by \$648 + \$178 = \$826 than it was in 1970. However, if the family's total expenditures were fixed in nominal terms, then an additional transfer of \$4745 − \$178 = \$4567 is required to put it in an equivalent economic position as in 1970. While this example is extreme, it illustrates the nonsense of a single-price focus. Losses experienced during 1970–80 due to price behavior should be viewed as problems of insufficient income for all goods, and are better treated as general poverty problems than problems of critical reliance on a particular commodity.

[15] Two points not explicitly considered are as follows. First, the results of this paper focus primarily on the welfare effects from consumption, without looking at indirect effect of price changes on the incomes of families, excepting the measures of Table 1 that correspond to total expenditure paths following per capita income growth rates. While there was some variation in regional average income over the decade, it is difficult to isolate how much of the variation is due to relative price shifts, as opposed to ordinary migration, regional unemployment changes, etc. Second, I have not explicitly isolated the induced effect of energy price increases on capital purchases via weatherization and other conservation measures, because the size of this effect appears relatively small (see Bernard Frieden and Kermit Baker, 1983, among others).

APPENDIX: DATA UTILIZED

	Energy	Food & Clothing	Other Nondurables	Capital Services	Consumer Services	ACPI	CPI
Nominal Prices							
1970	0.941	0.931	1.018	0.776	0.908	0.898	0.928
1971	0.972	0.959	1.073	0.807	0.955	0.937	0.968
1972	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1973	1.085	1.099	1.028	1.135	1.050	1.081	1.062
1974	1.407	1.227	1.310	1.038	1.129	1.170	1.178
1975	1.543	1.305	1.585	0.982	1.212	1.252	1.286
1976	1.645	1.342	1.642	1.163	1.284	1.346	1.360
1977	1.781	1.396	1.688	1.414	1.369	1.466	1.448
1978	1.887	1.501	1.824	1.548	1.477	1.586	1.559
1979	2.347	1.625	2.234	1.566	1.599	1.749	1.732
1980	3.019	1.737	2.904	1.496	1.746	1.951	1.969
Real Prices (1972 dollars)							
1970	1.047	1.036	1.133	0.864	1.011		
1971	1.037	1.023	1.144	0.860	1.018		
1972	1.0	1.0	1.0	1.0	1.0		
1973	1.003	1.016	0.951	1.049	0.971		
1974	1.202	1.048	1.119	0.887	0.964		
1975	1.232	1.042	1.266	0.784	0.968		
1976	1.222	0.997	1.220	0.864	0.953		
1977	1.214	0.952	1.151	0.964	0.933		
1978	1.189	0.946	1.150	0.976	0.931		
1979	1.341	0.929	1.277	0.895	0.914		
1980	1.547	0.890	1.488	0.766	0.894		

REFERENCES

Chipman, John S. and Moore, James C., "The Scope of Consumer's Surplus Arguments," in Anthony M. Tand et al., eds., *Evolution, Welfare and Time in Economics: Essays in Honour of Nicholas Georgescu-Roegen*, Lexington: Lexington-Heath, 1976, 69–123.

______ **and** ______, "Compensating Variation, Consumer's Surplus and Welfare," *American Economic Review*, December 1980, *70*, 933–49.

Christensen, Laurits R. and Jorgenson, Dale W., "U.S. Real Product and Real Factor Input, 1929–1967," *Review of Income and Wealth*, March 1970, *16*, 19–50.

Deaton, Angus S. and Muellbauer, John S., *Economics and Consumer Behavior*, Cambridge: Cambridge University Press, 1980.

Frieden, Bernard J. and Baker, Kermit, "The Market Needs Help: The Disappointing Record of Home Energy Conservation," *Journal of Policy Analysis and Management*, Spring 1983, *2*, 432–48.

Hausman, Jerry A., "Exact Consumer's Surplus and Deadweight Loss," *American Economic Review*, September 1981, *71*, 662–76.

Hicks, John R., "The Foundations of Welfare Economics," *Economic Journal*, December 1939, *49*, 696–712.

______, "Consumers' Surplus and Index Numbers," *Review of Economic Studies*, Summer 1942, *9*, 126–37.

Jorgenson, Dale W., "Econometric Methods for Modeling Producer Behavior," in Zvi Griliches and Michael D. Intriligator, eds., *Handbook of Econometrics*, Vol. III, New York: North-Holland, forthcoming 1986.

______, **Lau, Lawrence J. and Stoker, Thomas M.**, "Welfare Comparison under Exact Aggregation," *American Economic Review Proceedings*, May 1980, *70*, 268–72.

______, ______, **and** ______, "The Transcendental Logarithmic Model of Aggregate Consumer Behavior," in Robert Basmann and George Rhodes, eds., *Advances in Econometrics*, Vol. 1, Greenwich: JAI Press, 1982, 97–238.

______ and Slesnick, Daniel T., "Aggregate Consumer Behaviour and the Measurement of Inequality," *Review of Economic Studies*, July 1984, *51*, 369–92.

Muellbauer, John, (1974a) "Prices and Inequality: The United Kingdom Experience," *Economic Journal*, March 1974, *84*, 32–55.

______, (1974b) "Inequality Measures, Prices and Household Composition," *Review of Economic Studies*, October 1974, *41*, 493–504.

______, "Distributional Aspects of Price Comparisons," in Richard Stone and William Peterson, eds., *Econometric Contributions to Public Policy*, New York: St. Martin's Press, 1978.

Pollak, Robert A. and Wales, Terence J., "Welfare Comparisons and Equivalence Scales," *American Economic Review Proceedings*, May 1979, *69*, 216–21.

Samuelson, Paul A., "Complementarity—An Essay on the 40th Anniversary of the Hicks-Allen Revolution in Demand Theory," *Journal of Economic Literature*, December 1974, 12, 1255–89.

Stoker, Thomas M., "Bounds on Welfare Measures," in Robert Basmann and George Rhodes, eds., *Advances in Econometrics*, Vol. 4, Greenwich: JAI Press, 1985, 111–34.

Willig, Robert D., "Consumer's Surplus Without Apology," *American Economic Review*, September 1976, *66*, 589–97.

______, "Social Welfare Dominance," *American Economic Review Proceedings*, May 1981, *71*, 200–04.

______ and Bailey, Elizabeth E., "Income Distributional Concerns in Regulatory Policy-Making," in Gary Fromm, ed., *Economics of Public Regulation*, Cambridge: MIT Press, 1981, 79–118.

Terminating Hyperinflation in the Dismembered Habsburg Monarchy

By ELMUS WICKER*

The purpose of this paper is to measure the seriousness of the employment effects attending the ending of hyperinflation after World War I in the territories carved either completely or partially out of the Habsburg monarchy, notably Austria, Hungary, and Poland. Thomas Sargent (1983) has purported to show how hyperinflations were terminated dramatically in these newly created states without generating output and unemployment losses of the magnitude suggested by some recent estimates of Phillips curves. He acknowledged that disinflationary measures had the effect of substantially increasing unemployment, but the increase was small by comparison with dire predictions implied by certain estimates of the inflation-unemployment tradeoff.[1]

Sargent inferred from the abrupt ending of hyperinflation and the resulting magnitude of the employment changes that there was no "stubborn, self-sustaining momentum" preventing a rapid ending of inflation. He argued, moreover, that these alleged facts were consistent with the rational expectations hypothesis (*REH*). According to *REH*, future rates of inflation are conditioned by agents' perceptions of long-term government monetary and fiscal policies. Whenever there is a change in government strategy or policy, private economic agents can be expected to alter their strategies or rules for choosing consumption rates, investment rates, and portfolios. In other words, a change in the policy regime that is correctly perceived can terminate inflation without aggravating unemployment effects. Presumably, Austria, Hungary, and Poland underwent radical monetary and fiscal policy changes to end hyperinflation. Since hyperinflation ended abruptly without allegedly dire employment effects, the evidence, Sargent maintains, is at least consistent with the predictions of *REH*.

In assessing the merits of Sargent's hypothesis, a heavy burden falls on the interpretation of the unemployment data available in Austria, Hungary, and Poland. The absence of unemployment percentage estimates in Sargent's paper renders it virtually impossible to infer anything about the severity of unemployment. Nothing follows about the severity of unemployment by observing solely either the amount or changes in the amount of unemployment. I intend to remedy this shortcoming by estimating unemployment percentages following the introduction of monetary reform measures in each of the three territories.

My estimates, deliberately on the conservative side, show unemployment peaking in Poland after the introduction of measures to end hyperinflation at a little less than 13 percent; in Hungary at 12 percent and probably higher, and in Austria at 7 percent. The increased unemployment extended over a two to three-year interval; that is, before unemployment returned to its prereform levels. The behavior of the unemployment percentages demonstrates convincingly that the measures to end hyperinflation successfully in three central European states were at-

*Department of Economics, Indiana University, Bloomington, IN 47405. Helpful comments were generated by members of the Workshops in Economic History at the universities of Illinois and Indiana, especially Martin Spechler, Larry Neal, and Georgi Ranki. Bill Witte read and criticized several drafts. I owe a special debt of gratitude to Nicholas Spulber without whose expert knowledge of Eastern Europe I should never have undertaken a task that departs from my usual scholarly interests.

[1]Sargent refers to a "widely cited estimate," not otherwise identified, that for every one percentage point reduction in the annual inflation rate, $220 billion of annual *GNP* is lost (p. 42). It is indeed revealing that in summarizing what Sargent recognized as four important common features of the ending of four hyperinflations (including Germany), he failed to mention the behavior of unemployment.

tended by nonnegligible employment effects difficult to reconcile with *REH*. That the efforts to end hyperinflation did not generate dire output effects as predicted by recent estimates of the Phillips curve may be a valid criticism of contemporary estimates of the inflation-unemployment tradeoff, but it in no way detracts from the harshness and duration of the actual unemployment generated. Peter Garber also concluded, contrary to Sargent whose study included the German hyperinflation, that "there is substantial evidence of large-scale negative real effects in the aftermath of the German hyperinflation" (1982, p. 12).

Section I presents estimates of the unemployment percentage in Austria, Poland, and Hungary. Section II interprets the emergence of unemployment with the termination of hyperinflation.

TABLE 1—APPROXIMATE NUMBER OF UNEMPLOYED IN POLAND, 1922–26

	1922	1923	1924	1925	1926[c]
January	206[a]	81	101	176	360
February	[b]	107	111	185	
March	152	115	110	184	
April	128	113	95	178	
May	105	94	98	173	
June	87	76	138	172	
July	85	65	152	175	
August	69	57	165	186	
September	68	52	156	195	
October	61	56	147	214	
November	62	62	152	252	
December	75	68	162	311	

Source: Annuaire Statistique de la Republique Polonaise, Office Central De Statistique De La Republique Polonaise, 1923, 1924, 1925.
Note: Shown in thousands.
[a] End of month figures.
[b] Not available.
[c] Not available after January 1926.

I. Unemployment Percentage Estimates

A. *Termination of Hyperinflation in Poland*

Poland's first attempt to stabilize the zloty at a fixed parity with the dollar was a failure. But this attempt to end hyperinflation was an unequivocal success. Prices and exchange rates remained stable for eighteen months following efforts to halt hyperinflation in January 1924. It is true that this initial effort to peg the exchange rate was unsuccessful and that after July 1925, the inflation rate once again began to accelerate with the depreciation of the zloty. There was, however, no degeneration to a state of hyperinflation.

A series of drastic measures of monetary reform were inaugurated in January 1924 to halt hyperinflation. These included a capital levy, a progressive income tax, and the ending of all further note issues by the state. Prices and exchange rates abruptly stopped rising. The restoration of confidence was dramatic; price and exchange rate expectations were radically revised downward in response to the announcement of the monetary reform measures.

In May, the zloty-dollar exchange rate was fixed at $1 equal to 5.1826 zl. Difficulties began to emerge, however, in 1925. Budget receipts were far less than anticipated. Part of the explanation for the revenue deficiency was the decline in property tax yields due, perhaps, to postponement of tax payments the result of the previous year's poor harvest. Moreover, the proceeds of the capital levy fell far short of the original estimates. The budget deficit once again prompted the government to issue paper money. The zloty began to depreciate in the currency markets. The finance minister and head of state Ladislas Grabski resigned in November 1925. A financial crisis followed in the wake of his resignation—large deposit withdrawals accompanied by bank failures. Estimated total unemployment rose to over 360,000 by the beginning of 1926 (Table 1). At the end of May 1926, the exchange rate stood at 10.55 zl. to the dollar, more than double what it had been one year earlier. Wholesale prices had risen by as much as 50 percent between 1925 and 1926.

Two separate sets of official unemployment data exist for Poland in the post-World War I period: monthly estimates of the total number of unemployed since 1922 (Table 1) and monthly estimates of unemployed workers registered at State Employment Exchanges since 1924 (Table 2). There are no official statistics on the size of the total labor force in Poland between the two world wars.

TABLE 2—REGISTERED UNEMPLOYED IN POLAND, 1921–26

	1921	1922	1923	1924[a]	1924[b]	1925	1926
January	50	56	45	39	18	104	251
February	50	60	51	42	23	121	301
March	54	70	61	48	25	134	302
April	66	62	67	49	26	138	296
May	63	58	63	50	22	137	272
June	55	49	57	48	24	132	257
July	53	42	51	62	30	127	243
August	55	41	48	120	41	132	223
September	55	42	43	139	95	143	205
October	57	40	46	149	107	153	185
November	57	36	41	131	102	172	168
December	52	33	35	128	101	204	168

Source: Annuaire Statistique de la Republique Polonaise, 1920–29.
Note: Shown in thousands.
[a]Unemployment figures for 1921–24 refer to "in the course of the month."
[b]1924–26 refer to the beginning of the month.

Nevertheless, data do exist that will enable us to construct annual estimates of the unemployment percentage from 1922 to 1930. The numerator of the ratio can be derived from the number of workers registered at State Employment Exchanges and the denominator from the number of employed manual and mental workers insured under the compulsory health insurance scheme plus the registered unemployed. A brief description of the nature of the legislation creating employment exchanges as well as compulsory health insurance may be useful in interpreting the estimated unemployment percentages.

In January 1919, Poland established a system of State Employment Exchanges subordinate to a central employment service at the Ministry of Labor and Social Assistance. Employers were required to list all vacancies with the Exchanges. But the incentive for the unemployed to register did not become particularly strong until the introduction of compulsory unemployment insurance in August 1924. Since the Act provided that only unemployed persons registered at the State Employment Exchanges were entitled to benefits, there was a sizable increase in the percentage of registered to total unemployed; it increased from 22 percent in June 1924 to 69 percent in September 1924, gradually rising to 83 percent in December 1925. Compulsory unemployment insurance was extended to all workers 18 years of age and older employed in firms with more than 5 workers. Coverage included workers in industry, mining, metalworking, commercial undertakings, communications and transport. There was a further extension to some workers in state-supported enterprises in 1925. Intellectual or mental workers were excluded at first but were made eligible in 1926.

Compulsory sickness and maternity insurance was inaugurated in May 1920; the coverage was extensive. It included all manual and nonmanual workers, outdoor workers, and domestic servants whose wages did not exceed 125 zloty ($24) a week; civil servants, local government employees, and workers engaged in casual employment were excluded. Agricultural workers, except those in Upper Silesia, were included in a separate scheme. Contributions to support health care insurance were shared equally between workers and employers.

Column 2 in Table 3 shows the number of workers eligible for health insurance annually from 1923 to 1931 excluding Upper Silesia. Estimates of the number of workers having compulsory health insurance in Upper Silesia are shown in column 3.[2] Column

[2]Estimates for the number of workers having compulsory health insurance in Upper Silesia were obtained

TABLE 3—POLISH UNEMPLOYMENT PERCENTAGE ESTIMATES, 1922–30

Year[a] (1)	Compulsory Health Insurance: (Excluding Upper Silesia) (2)	Compulsory Health Insurance: Upper Silesia (3)	Total (2+3) (4)	Registered Unemployed (Including Upper Silesia) (5)	Total (4+5) (6)	Unemployment percentage (5/6) (7)
1922	1416	212	1628	20	1648	1.2
1923	1602	240	1842	18	1860	.96
1924	1653	248	1901	113	2014	5.6
1925	1602	240	1842	268	2110	12.7
1926	1783	267	2050	190	2240	8.5
1927	1991	299	2290	165	2455	6.7
1928	2277	342	2619	126	2745	4.6
1929	2265	339	2604	185	2789	6.6
1930	2172	338	2510	300	2810	10.7

Sources: Col. 2: *Concise Statistical Year-Book of Poland*, 1931, p. 111; Col. 3: Estimated for 1922–38 using actual observations on those with compulsory health insurance for Upper Silesia in 1929 and 1930 as a percentage of col. 2; (col. 2×15 percent). Quarterly data for Upper Silesia are contained in *Concise Statistical Year-Book of Poland*, 1930, p. 97; Col. 5: Table 2. Estimated for 1922. Registered unemployed does not include nonmanual or "mental workers" before 1926. Number of mental workers added to Registered Unemployed for period before 1926. Data for 1926–31: *Concise Statistical Year-Book of Poland*, 1934, p. 118.

[a]Data for 1922 refer to January 1, 1923: similarly for 1923–30.

5 shows the number of registered unemployed. Adding columns 4 and 5 provides the denominator of the unemployment ratio (col. 6). Column 7 sets out the estimates of the unemployment percentage for Poland. The figures for 1922 and 1923 grossly underestimate the unemployment percentage because of the low ratio of registered to total unemployed before the advent of compulsory unemployment insurance. Even if we assume that the registered unemployed were equal to total unemployment, the ratio would not exceed 3 percent in 1923. By assuming that the same ratio of registered to total unemployment prevailed in 1923 as in 1924, the unemployment percentage in 1923 would stand at 1.9 percent, the basis for the only viable comparison with the changes wrought in 1925 by the ending of hyperinflation since there are no comparable prewar unemployment data. The unemployment percentage had increased to 5.6 percent by the end of 1924, an increase of at least 3.5 percentage points from the preceding year; it rose to 12.7 percent in 1925 and did not fall below the 1924 level until 1928.

for 1922–28 by applying the percentage of observed number of insured workers in Upper Silesia in 1929 and 1930 to those compulsorily insured in 1929 and 1930 (excluding Upper Silesia) to figures in col. 2 (Table 3).

The facts then are quite clear; the unemployment effects attending the ending of hyperinflation in Poland were serious; moreover, they were not confined to a brief time interval. The full impact of monetary reform was delayed a year until 1925 with diminishing effects discernible in 1926 and 1927.

Both the numerator and the denominator of the unemployment ratio understate the amount of unemployment as well as the size of the labor force. There is no reason to think, however, that my estimates of the unemployment percentages are seriously biased either in an upward or downward direction, although the Report of the Bank of Poland for 1925 as quoted by Bernard Blumenstrauch (1932, p. 35) states that unemployment reached about 35 percent of the workers in industry. The incidence of unemployment among workers in industry may very well have been higher than the average. If my estimates err, it is more likely to be on the side of excessive caution rather than reckless optimism. They are probably the best that are attainable given the availability of Polish labor statistics.

TABLE 4—UNEMPLOYED IN RECEIPT OF BENEFITS IN AUSTRIA: MONTHLY, 1922–26

	1921[a]	1922	1923	1924	1925	1926
January		34	161	120	187	231
February		43	167	126	189	229
March	10	42	153	107	176	202
April		44	132	83	148	173
May		39	108	69	131	155
June	11	33	93	64	118	151
July		31	87	66	117	152
August		31	84	74	116	151
September	11	38	79	78	119	148
October		58	76	89	131	151
November		83	79	113	159	169
December	17	117	98	154	208	205

Source: Statistisches Handbook Fur Die Republic Osterreich, 1925, p. 120; 1927, p. 142.
Note: Shown in thousands.
[a]Data are only available quarterly for 1921.

Sargent inferred that the increase in unemployment in late 1924 was

> not in order of magnitude worse than before the stabilization, and certainly not anywhere nearly as bad as would be predicted by application of the same method of analysis that was used to fabricate the prediction for the contemporary United States that each percentage point reduction in inflation would require a reduction of $220 billion in real *GNP*. [p. 73]

But we have to go back two years or more to December 1921 and March 1922 to find comparable numbers of unemployed. That says nothing about the substantial increase —more than 3.5 percentage points—in the second half of 1924 coincident with the measures enacted to end hyperinflation. Moreover, Sargent failed to recognize that the consequences of attempting to stabilize the price level persisted into 1925 and 1926, thereby underestimating the total employment effects of terminating hyperinflation.

B. *Termination of Hyperinflation in Austria*

Austria is another example where hyperinflation was brought to a halt dramatically in September 1922 with the advent of monetary reform. But the employment effects were less severe than the Polish episode. Sargent recognized that there was a striking increase in unemployment, but, in the absence of unemployment percentages, he judged its harshness solely by contrast with what unemployment might have been if estimated by applying the 1 percent drop in inflation, $220 billion in lost output rule. We ought to be able to do better than that.

TABLE 5—REGISTERED UNEMPLOYED IN AUSTRIA MONTHLY, 1924–26

	1924	1925	1926
January	151	219	254
February	156	221	251
March	139	210	225
April	114	183	194
May	102	162	177
June	92	149	173
July	98	150	174
August	109	149	177
September	109	152	177
October	121	167	181
November	146	197	202
December	184	243	241

Source: Statistisches Handbook Fur Die Republic Osterreich, 1927, p. 142.
Note: Shown in thousands.

Continuous monthly data on unemployment in Austria are available since 1922 from the Unemployment Insurance Service. Two separate series are reported. From 1922 onwards data are reported on unemployed persons receiving relief payments (Table 4). Beginning in 1924 the registered unemployed as well as persons receiving unemployment benefits are reported separately (Table 5). The

TABLE 6—UNEMPLOYMENT PERCENTAGE AND REAL *GNP* ESTIMATES FOR AUSTRIA ANNUALLY, 1913 AND 1920–37

	Working Population (1)	Registered Unemployed (2)	Labor Force (1)+(2) (3)	Unemployment Percentage (4)	Real *GNP* (1937 Prices) (million) (5)
1913	3470	70	3540	2	10,802
1920					7,175
1921	–	–	–	–	7,942
1922	–	–	–	–	8,657
1923	–	–	–	–	8,562
1924	3298	188	3486	5.4	9,565
1925	3263	220	3483	6.3	10,211
1926	3235	244	3479	7.	10,378
1927	3260	217	3477	6.2	10,697
1928	3294	183	3477	5.3	11,194
1929	3282	192	3474	5.5	11,358
1930	3221	243	3464	7.	11,042
1931	3111	334	3445	9.7	10,154
1932	2959	468	3427	13.7	9,107
1933	2852	557	3409	16.3	8,803
1934	2845	545	3390	16.1	8,875
1935	2874	515	3389	15.2	9,056
1936	2874	515	3389	15.2	9,321
1937	2925	464	3389	13.7	9,822

Source: A. Kansel, N. Nerveth, and H. Seidel (1965, pp. 38 and 44). Martin Spechler brought this important reference to my attention.
Note: Cols. 1–3 shown in thousands.

Austrian Institute for Economic Research has published data on unemployment and the size of the labor force with which to compute unemployment percentages. These data for the years 1913 and 1924–37 appear in Table 6. An International Labour Office study (1925, p. 90) estimates unemployment in 1913 at 2 percent, but the average percentage was probably nearer 4 for the few years immediately preceding World War I (1910–14).

Unemployment began to accelerate from the time the Austrian Crown stablized in August 1922 to February 1923 (Table 4). We have no estimates of the unemployment percentage from 1914 to 1924. Nevertheless, the Austrian Trade Union Committee undertook to determine the percentage of totally and partially unemployed among some 600,000 members of various workers' organizations in December 1922; the results of the survey as revealed in an International Labour Office report showed that 33 percent were totally unemployed and 22 percent were partially unemployed (pp. 89–90). This large number of unemployed trade unionists is not reflected in Table 4—the unemployed in receipt of benefits—where the number of unemployed is stated to be 117,000. I do not know what the total unemployment percentage was in either 1922 or 1923, other than to say that it had begun to rise. According to Table 6 it reached a peak of 7 percent in 1926 and did not fall much below the level attained in 1924 thereafter. The increases in unemployment during the winters of 1924–25 and 1925–26 can be attributable largely, though not entirely, to the earlier measures of monetary reform.

These estimates are probably too low because all of the unemployed were not registered at the Employment Exchanges. Taking average unemployment to be at least 3 percent in 1921, the currency stabilization policies resulted in at least a 4 percentage point increase in the unemployment percentage, a not insignificant increase assignable to currency reform.

What is even more to the point, however, is the behavior of the real *GNP* as set out in column 5 of Table 6. Real *GNP* declined 1 percent in 1923 (currency reform intro-

duced in September 1922), but rose 10 percent in 1924, 7 percent in 1925, and 1.6 percent in 1926. The increase in unemployment was not mirrored in the behavior of output!

The Austrian case is the only one of the three countries where output figures are available. Even if Sargent underestimated the unemployment effects of terminating hyperinflation, he probably omitted inadvertently the positive effects of currency stabilization on output. Annual data, of course, conceal the output effects observable only on a quarterly basis. We do not know, for example, over how many quarters of 1923 and 1924 that stabilization exerted a depressing effect on output, or how depressing the effects were.

C. *Termination of Hyperinflation in Hungary*

Inflation was terminated in Hungary in June 1924 as it was in Austria earlier by monetary reform supervised and implemented by the Financial Committee of the League of Nations. The League proposed a loan to Hungary to cover the country's budget deficit until receipts and expenditures were once again equilibrated. Budget balance was achieved by July 1, 1924, and no further releases were necessary from the reconstruction loan. When the new central bank was opened on June 23, 1924, the Hungarian crown was stabilized in relation to the pound sterling (346,000 crowns to one English pound). In the year following stabilization, the gold reserve increased threefold. The retail price index fell 17 percent. Arthur Salter identified the less favorable factors attending the demise of inflation as a) "a considerable number of bankruptcies" and b) "some unemployment due to the process of adjustment to the new conditions of stabilization" (1926, pp. 29–30).

The effects of deflation are evident in the behavior of the amount of unemployment. The measurement of total unemployment in Hungary is complicated by the absence of official estimates except those derived from the Census of Population at the end of 1920. At that time, unemployment stood at 40,115 in Budapest and 65 other industrial towns.

TABLE 7—UNEMPLOYED TRADE UNIONISTS IN HUNGARY, 1922–25

Year	Unemployed Trade Unionists	Unemployment Democratic Socialist Union / Total Membership Democratic-Social Union
1922		
2 December	27,287	7.4
1923		
January[a]	–	
February[a]	–	
31 March	18,650	8.5
28 April	19,481	8.5
26 May	12,083	13.7
June[a]	–	
28 July	13,349	6.5
31 August	10,870	5.3
29 September	13,422	6.6
27 October	13,227	6.5
30 November	15,409	8.1
31 December	15,432	5.1
1924		
26 January	17,643	10.3
29 February	23,760	14.4
30 March	22,436	13.6
26 April	22,327	11.6
31 May[b]	23,412	11.7
28 June	25,574	12.8
26 July	30,004	15.
31 August	32,059	15.
27 September	29,377	13.5
31 October	32,299	15.7
29 November	33,028	16.
27 December	34,844	17.3
1925		
31 January	38,457	20.2
28 February	38,206	20.3
28 March	38,191	20.5
25 April	37,682	20.2
30 May	35,246	–
27 June	35,210	–
July	33,063	–
August	28,217	–
September	26,111	–
October	23,771	–
November	27,488	–

Source: Revue Hongroise De Statistique (1923–1925).
[a]Not available.
[b]Figures for Christian-Social Union are included.

Beginning in December 1922, however, we have continuous monthly estimates of the unemployed among trade union members. These data are available in *Revue Hongroise De Statistique* published by the Central Statistical Office. Trade union membership is

TABLE 8—TRADE UNION UNEMPLOYED, WORKERS WITH COMPULSORY HEALTH INSURANCE, AND UNEMPLOYMENT PERCENTAGE ESTIMATES: QUARTERLY, HUNGARY, 1924–26

	Number of Members Compulsorily Insured (1)	Observed Trade Union Unemployment (2)	Percent Unemployed (2/1) (3)	1.5× Col. 2 (4)	Percent Unemployed (4/1) (5)	2× Col. 2 (6)	Percent Unemployed (6/1) (7)	Percent 2.5× Col. 2 (8)	Unemployed (8/1) (9)
1924									
31 March	726,281	22,436	3.1	33,654	4.6	44,872	6.2	56,090	7.7
28 June	724,119	25,574	3.5	38,361	5.3	51,148	7.1	63,935	8.8
27 September	724,777	29,377	4.1	44,066	6.1	58,754	8.1	93,443	10.1
31 December	690,391	34,844	5.	52,266	7.6	69,688	10.1	87,110	12.6
1925									
31 March	659,123	38,191	5.8	57,287	8.7	76,382	11.6	95,478	14.5
30 June	661,012	35,210	5.3	52,815	8.	70,420	10.7	88,025	13.3
30 September	685,526	26,111	3.9	39,166	5.7	52,222	7.6	65,278	9.5
31 December	650,740	27,984	4.3	41,976	6.5	55,968	8.6	69,960	10.8
1926									
31 March	681,880	30,918	4.5	46,377	6.8	61,836	9.1	77,295	11.3
30 June	698,770	27,048	3.9	40,572	5.8	54,096	7.7	67,620	9.7
30 September	734,127	21,332	2.9	21,998	4.4	42,664	5.8	53,330	7.3
31 December	703,514	22,332	3.2	33,498	4.8	44,664	6.3	55,830	7.9

Source: Cols. 1 and 2: *Revue Hongroise De Statistique*, 1928.

confined to the Social-Democratic and Christian-Social unions. Membership in the larger of the two trade unions, the Social-Democratic, was 203,000 in 1922; according to Josef Vago, it had fallen to 179,000 in 1924 (1925, p. 347). These estimates cover workers who paid their union subscriptions but omit the unemployed and the sick who were exempt. Membership in the Christian-Social trade union probably did not exceed 16,500 in May 1924 when unemployment data for that union were first published. Table 7 shows monthly unemployment data for the Social-Democratic trade union from December 1922 through April 1924. After that date the figures include the unemployed members of the Christian-Social union as well. The last column shows the percentage of unemployed Social-Democratic trade unionists to total Social-Democratic membership from December 1922 to April 1925. Monthly membership data for the Christian-Social union were not available. The data reveal clearly a decisive and substantial increase in unemployment among trade unionists beginning in May 1924—both in terms of absolute numbers and percentages.

The economic significance of these data resides in the proportion of membership in trade unions to the total number of workers in commerce, mining, and industry. Since we do not know the actual number of workers in these activities, we must be content with a good proxy. Data on compulsory sickness and accident insurance exist since 1922 and may serve as an approximation to the total number of workers in commerce, industry, and mining although the figures do not include domestic servants and railway employees. Trade union membership including unemployed trade unionists constituted around 30 percent of those compulsorily insured in 1924. But it does not follow necessarily that total unemployment was $3\frac{1}{3}$ times the amount of trade union unemployment. Any estimate of the relationship between total unemployment and trade union unemployment is problematic, but a conservative estimate would suggest a multiple of less than 3 but greater than 1. I have arbitrarily selected 1.5 at the low end and 2.5 at the top, and constructed estimates of total unemployment quarterly from 1924 through 1926 (Table 8). From these new unemployment estimates, I have constructed unemployment percentages using quarterly data on the total number of workers subject to compulsory sickness and accident insurance.

The unemployment percentage rises from a low of 4.6 in March 1924 using the 1.5 multiplier, to a high of 8.7 percent in March 1925. It does not return to the 4.6 level again for two and one-half years. Employing a slightly more generous but still conservative multiplier of 2.5, the unemployment ratio rises from 7.7 percent in March 1924 to a high of 14.5 percent in March 1925.

It is not at all implausible to consider a multiplier of three, but I have deliberately preferred to err, if indeed I have erred, on the side of prudence and caution. However, raising the multiplier to three would raise the unemployment percentage from 9.2 in March 1924 to 17.4 percent in March 1925.

Another indicator of the behavior of unemployment in Hungary is available from the Office of Public Placement where annual data are recorded on employer job offers, demand for jobs, and placement (Table 9). The severity of the unemployment situation in 1925 and 1926 is indicated by the marked rise in the ratio of the demand for jobs (registered unemployed) relative to jobs available (employer job offers); that is, from 135 in 1924 to 181 in 1925. The increase in the ratio was due primarily to the 31 percent decline in employer job offers.

One advantage of the trade union data is that it is possible to measure the incidence of unemployment among various classifications of workers. Unemployment in the Democratic-Socialist union increased from 23,760 in February 1924 to 36,764 a year later, a net increase of 13,000. Table 10 shows approximately 80 percent of the net increase can be accounted for by the behavior of unemployment among four categories of workers. The most striking thing about these figures is the extraordinary increase in the number of unemployed in the financial sector—31.5 percent of the total net increase of 13,000. All of this increase can be attributable to the ending of hyperinflation which had increased substantially the money market as well as other operations of the commercial banks. The numbers of private banks, bill brokers, and commission agents had also increased in response to the acceleration of the rate of inflation. With the ending of hyperinflation and the subsequent reduction in the scale of bank operations, large numbers became redundant. In Budapest alone, Vago estimated that of the 12,000 bank officials, at least 4,000 were given termination notices in 1924 (p. 350). In Austria, perhaps as many as 10,000 bank employees lost their jobs.

TABLE 9—EMPLOYMENT OFFERS, DEMAND FOR JOBS, AND PLACEMENT

Year	Job Offers (1)	Demand for Jobs (2)	Job Placement (3)	Ratio of Col. 2 to Col. 1 (4)
1918	45	45	26	100
1919	97	67	34	69
1920	91	114	48	125
1921	87	116	55	138
1922	103	116	61	113
1923	80	105	52	131
1924	78	105	49	135
1925	54	98	39	181
1926	59	100	42	170

Source: *Revue Hongroise De Statistique* (1927, p. 753).
Note: Shown in thousands.

TABLE 10—INCIDENCE OF TRADE UNION UNEMPLOYMENT AMONG VARIOUS WORKER CLASSIFICATIONS: FEBRUARY 1924 TO FEBRUARY 1925

	Net Increase in Number Unemployed	Percent of Total Net Increase
1. Construction Workers	1926	14.8
2. Iron Workers	1726	13.3
3. Employees of Financial Establishments	4093	31.5
4. Woodworkers	2299	17.7
Total	10,044	77.3

Source: *Revue Hongroise De Statistique* (1925).

The impact on the building industry was likewise severe, but less so than on employees of the financial sector. Tight money policies engendered by the stabilization measures contributed to curtailing activity in the building industry including the slate and cement industries as well.

There is an additional source of data on unemployment in Hungary that derives from the Hungarian Employers Association which

TABLE 11—NUMBER OF WORKERS EMPLOYED IN ORGANIZED FACTORY INDUSTRIES 1923 AND 1924 IN HUNGARY

Industry	End 1923	1924 May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Coal Mining	55	55	55	55	50	49	47	46	45
Engineering	50	44	42	41	39	39	39	39	38
Ironware	3	3	2	2	2	2	2	2	2
Wood	10	9	8	7	7	7	7	7	7
Textiles	22	22	22	22	22	22	22	22	32
Chemical	22	19	15	14	14	14	14	14	14
Leather	5	5	5	5	5	5	5	5	5
Boots and Shoes	11	10	9	9	9	8	7	8	8
Printing	6	6	5	5	5	5	5	5	5
Bookbinding	2	2	2	2	1	1	2	2	2
Lithography	1	1	1	1	2	2	1	1	1
Brickworks	16	16	12	9	2	2	1	1	.4
Building	7	5	5	5	4	3	3	3	3
Glass	1	1	1	1	1	1	1	1	1
Milling (Budapest)	2	2	2	2	2	2	2	2	2
Total	213	200	186	180	164	162	158	158	165.4
Increase (+) or Decrease (−) from the Previous Month		−13	−14	−6	−16	−2	−4	0	+7.4

Source: Vago (p. 354).
Note: Shown in thousands.

began to publish figures in May 1924 on the number of workers employed by its members. There is a benchmark estimate for the end of 1923 as well (Table 11). The decrease in the number of employed workers is not a reliable indicator of the increase in unemployment because some of the discharged workers may have found employment in agriculture and other branches of industry. But the data do show that employment reductions were greatest in coal mining, engineering, chemicals, building, and brickworks. The sharp reductions in chemicals and the engineering industries are not reflected in the trade union figures probably because of the relatively small share of organized to unorganized workers in these industries. The trade union data (Table 10) show only 263 unemployed in the chemical industry whereas the Employers Association reveals a reduction of over 17,000. Similarly, the trade union data show little or no change in unemployment in the iron and engineering industry, but the Employers Association shows a decline of over 12,000.

II. Interpretations of the Unemployment Increase

In Section I, I attempted to show that the real effects of terminating hyperinflation in Austria, Hungary, and Poland were substantial as measured by the behavior of the unemployment percentage. I identified some especially strong influences that contributed to the emergence of unemployment in the wake of monetary reform. The reduction in the size of bank operations resulting from the ending of hyperinflation led to a sharp cutback in the number of employees of financial institutions in both Austria and Hungary—4,000 in Hungary, 10,000 in Austria. The attempt to curb budget deficits likewise led to a drastic reduction in the number of redundant government employees, 85,000 in Austria alone, and 38,000 in Hungary. Increased unemployment in the construction industry can be attributable in large part to high real rates of interest initiated by the newly established central banks as part of the monetary stabilization programs.

There were other considerations contributing to the increase in unemployment that arose directly out of the monetary reform measures. One of these was the alleged overvaluation of the Polish zloty in relation to the dollar.[3] The first series of monetary reforms in 1924 included a provision to define the dollar as equal to 5.18 zl. But six months after the currency reform, Polish prices began to rise at a faster pace than world prices thereby creating an import premium which was immediately reflected in an adverse trade balance.

Ragnar Nurske (1944) maintained that the principal cause of the disequilibrium was the unusually poor harvest of 1924; Polish crops in that year were nearly 33 percent less than normal. The rise in food imports necessitated by the crop failure aggravated the existing pressure on the balance of payments which led eventually in July 1925 to the fall of the zloty. According to Nurske, exchange stability could have been maintained "if Poland has [sic] possessed a monetary reserve adequate to bridge a transitory gap of this sort" (p. 26). An improvement in the harvest in the following year would have removed the source of the instability. To Feliks Mlynarski, Vice-President of the Bank of Poland, the fall of the zloty was unavoidable. Since Poland had been unsuccessful in floating an American loan, she was unable to cover the budget deficit generated by an unsatisfactory grain crop. Mlynarski wrote: "The stabilization of the currency in Poland—strange as it may seem—in July 1925 broke down because of a lack of 15 million dollars" (1926, p. 62).

Unlike Austria and Hungary, Poland undertook to stabilize her currency without intervention of the League of Nations. A foreign loan commitment was not, therefore, part of the initial Polish stabilization plan. The monetary reserves of the Bank of Poland were not adequate to withstand the temporary shock of the failure of the grain crop.

The effect of the fall of the zloty was a substantial contraction of imports. The gold value of monthly imports fell by about one-half in the second half of 1925—exports fell by one-fourth. In November the Premier Ladislas Grabski resigned. A financial crisis ensued. Twenty medium-sized and small banks failed or went into liquidation despite the efforts of the government to save them. Unemployment accelerated.

The interpretation of the increase in unemployment in Austria is not unambiguous. Some of the increase was undoubtedly due to delayed effects of major structural changes that had occurred when the Austro-Hungarian Empire was dismembered. One-third of Austria's 6 million inhabitants were concentrated in Vienna, once the banking, commercial, and administrative center for a nation of 29 million people. The huge bureaucratic apparatus centralized in the capital city of Vienna had become redundant. The elimination of superfluous civil servants and workers in state undertakings was inevitable though the process was necessarily accelerated by monetary and financial reform. The number of redundant government officials should not be underestimated. The Commissioner General's report shows that the Austrian government had agreed to a League of Nations proposal to dismiss 100,000 officials before July 1, 1924 (1923, pp. 4–5). The target, however, was not met. By that date only 68,331 had been discharged. Even at the time of the final report (1926) of the Commissioner General of the League for Austria (June 30, 1926), a total of 85,000 jobs had been terminated. Nevertheless, the number of dismissals was impressive and exacerbated unemployment following the advent of the stabilization measures.

Further contributing to the increase in unemployment was the dismissal of 10,000 bank employees who lost their jobs in 1924 due to the sharp contraction of bank operations with the terminaton of hyperinflation.

W. T. Layton and Charles Rist, who were invited by the League to make a comprehensive report on the situation of Austria, concluded that the real explanation for the increase in unemployment in the winter 1924–25 was the spur to efficiency engendered by

[3] Jack Taylor has argued that economic recovery was not possible as long as the zloty remained overvalued, but he provided no evidence to support his contention (1952, p. 39).

the ending of hyperinflation. They reasoned that business firms now had the inducement to become more cost efficient since they could not simply pass an increase on in the form of higher prices. They concluded: "Reorganization in the factories is thus the chief explanation of the apparent anomaly that unemployment has increased side by side with steady improvement in economic activity" (1925, pp. 16–17).

The rise in unemployment in Hungary, we attempted to show, was associated with sharp reductions in the scale of bank operations and activity in the iron and building trades induced primarily by stabilization policies.

Some of these influences, though not all, can be labelled longer-run structural considerations that came "home to roost" only after hyperinflation had ended. For example, the dismemberment of the Austro-Hungarian Empire drastically reduced the size of domestic markets. The population of the new Hungarian state was 8 million compared to 20 million in the old. The reduction in the number of government employees in both Austria and Hungary, though timed to coincide with the removal of budget deficits and monetary stabilization, can be attributed to dismemberment. Moreover, the abolition of the Customs Union within the old Habsburg monarchy and the resurgence of the imposition of tariffs changed radically the structure of trading relationships among the successor states. It would not be easy to ferret out the degree to which each of these factors affected poststabilization unemployment.

In addition to what I have labelled structural and monetary reform explanations for the increase in unemployment, two additional factors deserve some consideration: the role of unemployment insurance benefits, and the effects of the ending of the German hyperinflation on the contiguous economies of Austria, Poland, and Hungary.

Among the many factors affecting the decision to work is the amount of unemployment insurance benefits available. The more liberal the benefit payments, the greater the time and effort devoted to job search. Moreover, if unemployment insurance raises the benefits of unemployment above the value of the workers' marginal product, the employer may be induced to increase the number of layoffs rather than vary the wage. Daniel Benjamin and Lewis Kochin (1979) have argued persuasively that levels of unemployment in interwar Great Britain can be explained by the ratio of unemployment insurance benefits to wages and the continual liberal extension of eligibility requirements. Benefit-wage ratios averaged above 50 percent for much of the interwar period (1923–38). They emphasize, however, three special features of the British system that were conducive to lowering the level of desired unemployment: 1) worker insurance contributions were unrelated to past unemployment experience; 2) benefits were not tied to wages; and 3) benefits were payable for unemployment spells as short as one day (see pp. 446–47).

Government unemployment insurance schemes existed only in Austria and Poland. In Hungary trade unions paid some indeterminant amount of benefits to their unemployed members. Unemployment insurance in Austria applied to all workers except those engaged in agriculture, forestry, domestic service, and industry in certain rural districts. Benefits beginning in the eighth day of unemployment were granted originally for a period not exceeding 12 weeks per annum for unemployed persons who had worked at least 20 weeks during the preceding two years. But heavy unemployment in 1924 and 1925 led to an extension of the period to a maximum of 42 weeks. Benefits were stipulated in the form of daily allowances: unmarried workers received a flat rate of 2 schillings a day up to 3.1 schillings per day for married workers with three or more children. The average amount of the benefit was as low as 2.21 schillings per day. Since payments were not expressed as a percentage of wages, it is more difficult in the case of Austria to calculate an average unemployment benefit-wage ratio. Estimates indicate that highly skilled workers received an average wage of 45 to 50 schillings a week, skilled assistants 35 schillings, and women between 15 and 20 schillings. There are no estimates for unskilled workmen. Assuming the benefit payment for unmarried skilled workers is 14 schillings a week, that would imply a benefit-wage ratio

of 28 percent and for unmarried women a ratio between 70 and 90 percent! Beginning in September 1925 insurance recipients could receive no more than 80 percent of their weekly wage when fully employed. Layton and Rist explain the introduction of the benefit ceiling: "We understand that this limitation was introduced in order to prevent benefits discouraging the unemployed from seeking work in the open market. The percentage difference between the ruling wages on the market and unemployment benefits is smaller in Austria than in Germany and Poland" (p. 105).

Willingness to work was tested by requiring that the unemployed appear at least twice a week at the labor exchange. Exchange officials had the right to send the unemployed to fill vacant positions. The administration of the insurance fund was in the hands of industrial commissioners. In Vienna, for example, there were 50 comptrollers each of whom visited between 18 and 20 unemployed workmen a day. Layton and Rist (p. 105) concluded that in Vienna and large industrial districts, surveillance of the unemployed was good, unlike smaller country districts where it was less efficient.

Unemployment insurance in Poland was introduced in September 1924. Benefits for unemployed single workers amounted to 30 percent of wages up to a maximum of 5 zloty a day. Unlike Austria where the benefits were computed on a flat-rate basis, benefits were calculated as a percentage of total wages and graduated by the number of dependents. Single unemployed workers received 30 percent of wages up to a maximum of 5 zloty a day and as much as 50 percent for workers with 5 dependents. A single unemployed worker received no more than 1.5 zloty a day. Benefits were payable after the tenth day from the date of registration at the unemployment exchange and extending for not more than 13 weeks in each year. To qualify, the recipient must have worked for at least 20 weeks during the year preceding the unemployment spell.

Since there was no government unemployment in Hungary, it cannot explain the behavior of unemployment. And the Polish scheme was neither generous enough in terms of benefit payments nor length of coverage to have had a perceptible influence on the behavior of unemployment. The 30 percent benefit-wage ratio was close to Benjamin and Kochin's zero-effect level for Great Britain in the interwar period; they estimated that a benefit-wage ratio of .27 or below had no influence on desired employment. The situation in Austria was quite different. As in Great Britain, benefits were computed on a flat-rate basis, that is, not tied to wages. And the benefit-wage ratio was very generous indeed, especially for the lower paid workers. In Britain the maximum duration of coverage was 26 weeks during the heaviest unemployment spells between October 1931 and late 1934, whereas in Austria in 1924–25 the maximum period was 42 weeks. The argument that unemployment insurance might have affected worker preference for employment is at least plausible in the case of Austria, but any firm conclusions must await further empirical testing.

Another consideration that conceivably might have affected levels of unemployment in Austria, Hungary, and Poland was the successful ending of the German hyperinflation in November 1923. The massive shifts in demand across industries associated with the termination of German hyperinflation could easily have had repercussions on the volume and structure of trade in the contiguous territories. Garber attributed the increase in German unemployment to a deliberate policy of subsidizing the capital goods industry indirectly through inflation tax revenues, thereby changing the relative price of goods and factors of production.

There is no evidence that the ending of German hyperinflation had adverse effects on employment in Hungary in 1924 and 1925. Germany took less than 10 percent of Hungary's exports. Exports to Germany increased from 45 million gold crowns in 1924 to 69.4 million gold crowns in 1925, and the share of these exports to total exports increased from 7.9 percent to 9.92 percent during the two-year period. Imports from Germany also increased, rising from 88 million gold crowns in 1924 to 111 million gold crowns in 1925. The share in total imports rose from 12.5 to 15 percent.

There were, however, clearly discernible effects in Austria and Poland stemming from Germany's determined efforts in 1925 to improve her balance of payments position by waging commercial war on Poland. Germany prohibited the importation of Polish coal effective in June. The impact on Polish exports to Germany was immediate, falling from 57.7 million zloty in June to 32 million zloty in July and averaging approximately 37 million zloty for the remainder of the year. In addition to coal, the iron and zinc industries in Poland were directly affected as well. The Polish market was also important for German manufactured goods. Without having attempted to measure the effects quantitatively, Taylor concluded: "The various measures and countermeasures thus caused economic hardship on both sides" (p. 118).

Nor is there any direct evidence that Austria was adversely affected by the ending of German hyperinflation. Austrian exports to Germany increased 13 percent in the two years immediately following stabilization of the mark. But there were identifiable indirect effects stemming in part from Germany's embargo on Polish coal and the role, albeit a minor one, it might have played in contributing to the final breakdown in Polish stabilization in July 1925. Polish imports in the second half of 1925 had declined to the level of the year 1923. Austria was one of Poland's main trading partners. Austrian exports to Poland were more than halved between 1924 and 1926, having fallen from 194 million schillings to 73 million schillings in 1926.

III. Summary

To assess the validity of Sargent's claim that hyperinflation can be terminated without dire employment effects, I constructed monthly unemployment percentage estimates for Poland, Hungary, and Austria over the period both before and after monetary stabilization. These data reveal that there were significant effects on employment in the three territories, but no strong observable negative effects on output in Austria. My conclusions agree with those of Garber for Germany. These unemployment effects are difficult to reconcile with the rational expectations hypothesis. Admittedly, the effects on employment were less severe than what might have been predicted on the basis of modern estimates of the Phillips curve. But that in no way obscures the seriousness of the measured employment effects. Furthermore, one can question the appropriateness of using contemporary estimates of the Phillips curve to predict the behavior of employment and inflation in three newly established countries after World War I whose industrial structure bore no resemblance to that of the United States. Each of the three successor states was primarily an agrarian economy. At least 50 percent or more of the gainfully employed population was engaged in agriculture, and 20 percent or less in industry in both Hungary and Poland. About the same ratio prevailed in agriculture in Austria if forestry and kindred occupations are included. This salient fact cannot be ignored in assessing the impact on industrial unemployment of efforts to end hyperinflation.

I attempted to identify those considerations that contributed to the emergence of unemployment with the advent of monetary stabilization. Four different kinds of influence were isolated: 1) longer-term structural factors that came "home to roost" only after hyperinflation had ended, 2) shorter-term measures related more directly to monetary stabilization, 3) unemployment insurance benefits, and 4) delayed impact of the ending of German hyperinflation. In the first group were included unemployment attributable to reductions in the civil service timed to coincide with the removal of budget deficits and stabilization and to reductions in the number of employees of financial institutions with the termination of hyperinflation. The second group included alleged overvaluation of the Polish zloty and the Hungarian pengo, and policies to restrict credit initiated by the newly established central banks. The third and fourth represent special considerations not related necessarily to measures to stabilize the currency. Neither unemployment insurance benefits nor the ending of German hyperinflation can explain a significant amount of the increase in unemployment in Austria, Poland, and Hungary.

The increase in unemployment attending stabilization in Poland, Hungary, if not Austria, does not therefore seem to be consistent with the predictions of the rational expectations hypothesis.

REFERENCES

Benjamin, Daniel K. and Kochin, Lewis A., "Searching for an Explanation of Unemployment in Interwar Britain," *Journal of Political Economy*, June 1979, *87*, 441–78.

Blumenstrauch, Bernard, *Le Noveau regime monetaire en Pologne et son role dans l'economie nationale*, Nancy: Societe d'Impressions Typographiques, 1932.

Garber, Peter M., "Transition from Inflation to Price Stability," *Carnegie-Rochester Conference Series on Public Policy*, Vol. 16, *Regimes and Protection*, Autumn 1982, 11–42.

Kansel, A., Nerveth, N. and Seidel, H., *Osterreichs Volkseinkommer 1913 bis 1963*, Wien: In Selbstverlag Des Osterreichischen Institutes Fur Wirtschaftsforschung, 1965.

Layton, W. T. and Rist, Charles, *The Economic Situation of Austria*, Geneva: League of Nations, 1925.

Mlynarski, Feliks, *The International Significance of the Depreciation of the Zloty in 1925*, Warsaw: The Polish Economist, 1926.

Nurske, Ragnar, *International Currency Experience*, Geneva: League of Nations, 1944.

Salter, Arthur, "General Survey," *The Financial Reconstruction of Hungary*, Geneva: League of Nations, December 1926, 9–45.

Sargent, Thomas J., "The Ends of Four Big Inflations," in Robert Hall, ed., *Inflation: Causes and Effects*, Chicago: University of Chicago Press, 1983.

Taylor, J., *The Economic Development of Poland 1919–1950*, New York: Cornell University, Press, 1952.

Vago, Josef, "Unemployment in Hungary: Its Causes and Cure," *International Labour Review*, September 1925, *12*, 346–65.

Chief Bureau of Statistics of the Polish Republic, *Concise Statistical Year-Book of Poland*, Warsaw: Chief Bureau of Statistics, 1930, 1931, 1934.

Commissioner General of the League of Nations, *Financial Reconstruction of Austria*, First Report, Geneva: League of Nations, 1923; Final Report, 1926.

Office Central Des Statistique De La Republique Polonaise, *Annuaire Statistique de la Republique Polonaire*, Varsovie, 1920–29.

Revue Hongroise De Statistique, Redizee et Publiee Par L'Office Central De Statistique De Royaume De Hongrie, Budapest: 1923, 1925, 1927.

Statistiches Handbuch Fur Die Republic Osterreich, Bundesaupt Fur Statistik, Wien: Bundesomt Fur Statistick in Kommission Bei Karl Gerolds Sohn, 1925, 1927.

Unemployment Insurance and Labor Contracts under Asymmetric Information: Theory and Facts

By ANDREW J. OSWALD*

Over the last few years there has been a minor revolution in the theory of labor contracts. The early work of Costas Azariadis (1975), Martin Baily (1974), and Donald Gordon (1974) made the assumption that output prices, although uncertain *ex ante*, could be observed by workers *ex post*. Later work, however, has taken a rather different tack: Guillermo Calvo and Edmund Phelps (1977), Sanford Grossman and Oliver Hart (1981), Jerry Green and Charles Kahn (1983), and others have assumed that workers cannot observe the values of the uncertain variables.[1] Models of this type, where there is a distortion created by the existence of asymmetric information, turn out to have interesting properties. In particular, underemployment equilibria emerge.[2]

This is potentially important, because many economists see the explanation of unemployment as an especially desirable goal. However, as Hart's (1983) introductory survey points out, the analysis does not explain involuntary unemployment in the sense of identical workers being treated differently. There is never an unemployed individual who would like to swap places with an employed person with the same characteristics. Instead, the models explain how equilibria might arise in which there is less employment than in a competitive spot labor market. Although this is an achievement of some importance, it is not necessarily a solution to the main problem. Some economists take that to be the question of why in the real world there are workers who seem and claim to be involuntarily unemployed. In fact, that was what implicit contract theory managed to do fairly successfully: the models of Azariadis, Baily, and Gordon could generate equilibria in which unemployed individuals had lower utility than those workers with jobs. Unfortunately, those analyses also predicted more employment than under a competitive labor market, which was eventually seen as a serious flaw. George Akerlof and Hajime Miyazaki (1980) contains a well-known critique.

The purpose of this paper is to show that it is possible to combine traditional implicit contract theory with the assumption of asymmetric information. It is surprising that this has not been done before in the literature, because when one takes the two approaches together, it turns out that the model which emerges is capable of explaining the simultaneous occurrence of involuntary unemployment and underemployment. The literature has, of course, come very close to producing a combination of the old and new ideas. But one thing has been dropped from the recently developed theories of labor markets under asymmetric information—the assumption that workers have to rely on the government, or some other sector, for their income while unemployed. Grossman and Hart and others typically assume that firms pay unemployment insurance to those who do not get jobs at the prevailing output price. It is well known that the results of efficient contract models of all kinds are very sensitive to this type of insurance assumption.

*Centre for Labour Economics, London School of Economics, Houghton Street, London WC2A 2AE, England. I thank the Economics Department and the Industrial Relations Section at Princeton University for their warm hospitality during the academic year 1983–84. Helpful suggestions were made by David Card, Oliver Hart, Edmund Phelps, and the participants in a workshop at Nuffield College, Oxford.

[1]A collection of papers can be found in the 1983 *Quarterly Journal of Economics* Symposium issue. Robert Hall and David Lilien (1979) is another closely related paper.

[2]Overemployment results are also possible. See Green and Kahn, for example, for a discussion of the importance of the normality of leisure in an individual's utility function. This paper does not use the one-worker model which dominates the Symposium issue.

The introduction of an internal insurance scheme into conventional contract theory, for example, makes the model's equilibrium collapse to that of a competitive spot market.[3]

Should one assume that firms and workers run their own unemployment insurance schemes? The right answer to this is not immediately obvious. One approach—that implicit in Grossman and Hart, say—is to use the assumption on the grounds that under the other assumptions of the model it is rational for the firm and the workers to agree on this form of unemployment benefit plan. This is an attractive argument. The alternative approach, which ignores private unemployment insurance, also has some arguments in its favor. Bengt Holmstrom (1983), for example, claims that real world insurance schemes do not conform to the predictions made by the theory,[4] and argues that there is a moral hazard explanation: those workers without jobs would not have the appropriate incentives to find work. Holmstrom suggests that the analytical structure would be largely unaltered by the explicit introduction of moral hazard, which is why he omits all forms of unemployment pay from the analysis. Early writers on the topic took just this view; recent writers have tended to adopt the former position.

It is rather natural to try to settle this question by looking at what happens in the world. This is what the next section does. It studies data on American labor contracts which give us information about the extent of Supplemental Unemployment Benefit (SUB) plans and severance pay. The contracts are for the years 1970, 1975, and 1980, and cover about six million unionized employees. The very broad conclusion is that the extent of SUB plans is rather greater in manufacturing industry than may well be realized (although about half the workers still do not have such a scheme), but that SUB plans are almost nonexistent in the nonmanufacturing sector. This suggests that in that sector, and a large part of manufacturing industry, it seems to be sensible to assume that there is no private unemployment insurance. Section III pursues this idea. It sets out a model of a labor contract under asymmetric information in which workers cannot rely on the firm to make unemployment benefit payments. The analysis blends the assumptions of early implicit contract theory with those of modern contract models.

If private unemployment insurance is rare in the world, there is likely to be a good reason. Section IV develops this point. It sets out a model in which there are layoffs by seniority (a feature of almost all U.S. labor contracts) rather than, as is implicit in most conventional theory, unemployment by random draw. The model solves for an efficient contract (hence takes a quite different form from, for example, the seniority model in Gene Grossman, 1983).[5] Its central ideas are that the seniority ordering insulates the majority of workers from product market fluctuations, and that this reduces the incentive for a labor group to vote for a supplemental unemployment benefit plan.

I. The Extent of Unemployment Insurance Schemes in the United States

The first important Supplemental Unemployment Benefit plan in the United States was negotiated in the automobile industry, between Ford and the UAW, in 1955. Others quickly followed. By 1970, about two million workers were covered in this way, and the figures given below suggest that coverage in union establishments has grown since then. The aim of the SUB plans is to compensate temporarily unemployed workers more fully than is done by government payments alone. More than a decade ago, E. F. Beal et al. (1972) stated that plans did not generally make up unemployed income levels to more than 75 percent of normal earnings. Today, at least some are more generous. The 1982 labor contract signed by the UAW and General Motors has provision, subject to certain

[3] See Grossman and Hart, for example.

[4] Holmstrom presents no evidence, but his view is largely confirmed by the data given in the next section.

[5] In Grossman's model, the union maximizes against the labor demand curve.

caveats, for "an amount which, when added to...State Benefit and Other Compensation, will equal 95% of...Weekly After-Tax Pay, minus $12.50, to take into account work-related expenses not incurred." Similarly, the 1982 agreement between the Goodyear Tire Company and the URCLPWA fixed an amount normally equal to 80 percent of a worker's pay, or, under special conditions, 100 percent of pay in the event of a factory closure. It would obviously be desirable to obtain systematic evidence on the generosity of private unemployment insurance provisions in U.S. labor contracts. Unfortunately it appears that there are no data of this kind. Hence we are forced to rely on these relatively casual observations about particular labor contracts.

Some firms also have severance pay schemes. This payment is made only to those who are permanently separated from the employer,[6] and it usually requires a minimum length of service. The 1982 contract agreed upon between Air Canada and the IAMAW, for example, states that employees with five years' service must be paid 5 weeks' pay as a severance allowance. The maximum is 20 weeks' pay for twenty years or more of employment.[7] It is worth stressing that this type of insurance is not obviously of the form typically assumed in the theoretical work on labor contracts. There, unemployment insurance is usually to be thought of as money paid intermittently to those in the labor pool. Severance pay in the real world is a method of altering the size of the future labor pool. The final sort of job insurance specified in American contracts is the so-called wage-employment guarantee, which provides income for twelve months of the year regardless of hours worked. As the data below show, however, these are still fairly uncommon.

[6]This is an important issue in contract negotiations whenever significant technical change is expected. The 1983 agreement between New Jersey Bell and the International Brotherhood of Electrical Workers is a good example.

[7]The four agreements mentioned in this section were drawn at random, and are unlikely to be exceptional, but without a proper sample it is not possible to be certain.

There is a rich data source—although it does not seem to be as widely known as might be expected—on the extent of these unemployment insurance arrangements. This source, *Characteristics of Major Collective Bargaining Agreements*, is available from the U.S. Bureau of Labor Statistics. Regrettably, however, the series has been discontinued, so the data reported here are some of the last which will be available. All the collective agreements covered in the samples involve establishments of 5000 workers or more in 1970, and 1000 workers or more in other years, and all are unionized plants and companies. Three of the most recently available years are presented below. Sample size in the tables refers to the number of labor contracts. Worker coverage was (in numbers of workers):

	Manufacturing	Nonmanufacturing
1970	2,362,075	1,741,050
1975	3,750,950	3,318,800
1980	3,025,150	3,568,650

Tables 1 and 2 give the figures on SUB plans and severance pay agreements. Table 3 has an industrial breakdown of SUB coverage for the most recent year, 1980. Table 4 contains data on the prevalence of wage-employment guarantees.

The tables speak for themselves, so it is only necessary to stress the main points.

1) Approximately half the workers in manufacturing industry had a SUB plan. In nonmanufacturing, the figure was less than 5 percent.

2) Approximately half the workers in manufacturing could claim severance pay if permanently laid off. About one-quarter of nonmanufacturing employees had this kind of compensation scheme.

3) Table 3 shows SUB and severance pay coverage by industry for 1980, and it illustrates the large differences in the extent of coverage. In some industries, such as Rubber and Plastics, three-quarters of the workers were covered by each type of plan; in others, like Construction, unemployment insurance was almost nonexistent.

4) Wage-employment guarantees, explained in Table 4, applied to roughly one in

TABLE 1—SUPPLEMENTAL UNEMPLOYMENT BENEFIT PLANS

		Percent of Agreements		Percent of Workers	
Year	Sample Size	Manufacturing	Nonmanufacturing	Manufacturing	Nonmanufacturing
1970	252	48.4	0.8	67.2	0.3
1975	1514	25.4	4.0	49.8	2.8
1980	1550	24.7	3.9	51.4	4.4

TABLE 2—SEVERANCE PAY PLANS

		Percent of Agreements		Percent of Workers	
Year	Sample Size	Manufacturing	Nonmanufacturing	Manufacturing	Nonmanufacturing
1970	252	52.4	31.7	62.8	34.5
1975	1514	41.7	20.0	51.6	22.3
1980	1550	43.2	24.9	53.6	27.0

TABLE 3—SUPPLEMENTAL UNEMPLOYMENT BENEFIT (SUB) PLANS AND SEVERANCE PAY BY INDUSTRY, 1980[a]

	SUB Plan		Severance Pay	
Industry	Percent of Agreements	Percent of Workers	Percent of Agreements	Percent of Workers
Manufacturing:				
Food, Kindred Products	3.8	3.0	51.9	67.4
Tobacco Manufacturing	25.0	21.5	100.0	100.0
Textile Mill Products	–	–	–	–
Apparel	48.4	41.7	6.4	3.6
Lumber, Wood Products	9.0	17.5	9.0	7.0
Furniture, Fixtures	11.8	13.4	11.8	9.1
Paper, Allied Products	–	–	47.6	53.4
Printing and Publishing	20.0	42.1	53.3	42.4
Chemicals	–	–	69.4	73.2
Petroleum Refining	–	–	60.0	64.5
Rubber and Plastics	64.3	82.6	50.0	75.2
Leather Products	–	–	54.5	70.3
Stone, Clay, and Glass	11.4	6.1	60.0	74.0
Primary Metals	72.7	9.1	59.0	81.5
Fabricated Metals	24.4	35.7	43.99	43.1
Nonelectrical Machinery	32.1	57.7	29.36	27.3
Electrical Machinery	9.6	21.3	48.2	55.9
Transportation Equipment	33.9	74.3	28.6	52.7
Instruments	–	–	36.4	30.4
Miscellaneous	–	–	44.0	50.0
Nonmanufacturing:				
Mining	25.0	6.6	31.2	7.3
Petroleum & Natural Gas:				
Transportation	–	–	11.3	6.4
Communications	1.25	1.7	91.2	89.3
Utilities, Electric & Gas	1.23	6.6	38.3	34.3
Wholesale Trade	8.3	4.4	25.0	20.7
Retail Trade	2.4	16.7	40.6	34.3
Hotel & Restaurants	–	–	6.4	2.4
Services	1.5	0.6	37.9	41.6
Construction	6.1	4.4	0.9	1.0
Miscellaneous	–	–	–	–

[a]Sample sizes for all Agreements and Workers were as follows: Agreements: 1,550; Workers: 6,593,800.

TABLE 4—WAGE EMPLOYMENT GUARANTEES

Year	Sample Size	Percent of Agreements		Percent of Workers	
		Nonmanufacturing	Nonmanufacturing	Manufacturing	Nonmanufacturing
1970	252	9.5	11.9	10.8	15.2
1975	1514	7.6	17.6	8.2	26.2
1980	1550	7.9	14.2	13.4	19.7

ten manufacturing workers and two in ten of the nonmanufacturing workers.

The extent of SUB plans is the most interesting question for this paper, because it comes closest to the spirit of the literature on labor contracts under asymmetric information. Severance pay coverage, although greater than SUB coverage, is less important here—partly because the size of payments tends to be surprisingly small (a week's pay, say, for each year of service), and partly because it does not correspond as closely to the notion of intermittent unemployment compensation paid by a single long-term employer. It may be best to think of severance pay more as insurance against losses in the value of human capital caused by market and technological changes. The issues do not seem to have been explored in the published literature. Wage-employment guarantees are intriguing, and reminiscent of traditional contract theory's wage rigidity result, but coverage is low.

It seems reasonable to draw the following broad conclusions. First, there are almost no nonmanufacturing workers who are covered by unemployment insurance schemes (like SUB plans) of the sort specified in the recent literature on labor contracts with asymmetric information. Even if we take the most flexible view of what an unemployment insurance arrangement might be, and add in severance pay and wage-employment guarantee schemes, only a minority of these workers appear to have cover against layoff or redundancy. Second, around half the workers in manufacturing have no unemployment insurance in a form like a SUB plan (or a wage-employment guarantee). These facts suggest that, if the object is to understand labor contracts of the type negotiated by American trade unions, it may be useful to study models in which the only unemployment benefit available to a worker is that statutory amount paid by local or federal government. One objection to this is that only 20 percent of U.S. workers are unionized and that contract theory was designed to explain nonunion labor markets. Natural counterarguments are that, first, the BLS data are the best we have about any large sample of employment contracts; second, union agreements might reasonably be expected to be if anything more generous (in unemployment insurance and other provisions) than those in nonunion sectors; and, third, the qualitative characteristics of union and nonunion labor markets are likely to be the same.

II. The Model

This section sets out a model of a labor contract under asymmetric information. The key assumption is that the firm does not pay unemployment insurance to those of its workers whom it lays off. We shall need a number of assumptions and it is sensible to keep them as conventional as possible.

1) Each worker has a utility function, $u(\cdot)$, defined on the open interval $(0,\infty)$. The function is concave, increasing, bounded, and twice differentiable. It is to be thought of as the utility from real income. Leisure is taken to have no value.[8] The wage income when working is w, and is b when not working. Only a single unit of labor is supplied by each worker.

2) There is a fixed labor pool, normalized at size unity. Employment is n, defined on the closed interval $[0,1]$.

3) The preferences of the labor group (or trade union) are described *ex post* by the

[8] This is inessential.

utility function $U = nu(w)+(1-n)u(b)$, which is a quasi-concave function defined on wages and employment.[9] Labor's preferences *ex ante* are described by the group's expected utility function, EU.

4) The firm or employer has a utility function, $v(\cdot)$, defined on the real line. It is concave, increasing, bounded, and twice differentiable, and is to be thought of as measuring the utility from profit income, π. The firm maximizes expected utility, Ev.

These assumptions give the nature of the agents' preferences. They are similar to assumptions made in the early contract literature and in Grossman and Hart, and Holmstrom inter alia. Another especially important assumption is the following:

5) The utility of an unemployed worker, $u(b)$, is independent of the firm's actions.

This is a key point in the paper. It follows Baily, Azariadis (1975), and Holmstrom, and plays a crucial role, as is now well known, in labor contract models under symmetric information.

It is also necessary to specify the form and nature of uncertainty, the structure of technology, and the minimum profit (utility) necessary to keep firms (workers) from withdrawing from the market.

6) The firm's output price, θ, is uncertain. It is distributed according to the density function $g(\theta)>0$ on closed support $[\underline{\theta},\bar{\theta}]$. Only the firm observes θ.

7) Output is given by a production function, $f(\cdot)$, defined on the unit interval. The function is concave, increasing, bounded, and twice differentiable.

8) A firm elsewhere in the economy earns expected utility v^*.

9) No one will work for the firm in state θ if $u(w)<u(b)$.

These assumptions describe a world of the following kind. A firm aims to employ some workers to produce a good which it will eventually sell in a competitive market. No one can be sure what the price will be, but the employer must recruit today. It would like to offer its workers a deal in which their pay depends on the selling price of the product. This cannot be done, however, because the employees will not be able to observe the price. (By way of example, auto workers, say, can observe for certain only the sticker prices on the automobiles they help to manufacture, and cannot know what dealer and customer discounts are being offered across the country.) To get any workers at all, it is assumed that the company must pay a wage that is greater than government unemployment benefit or the alternative reservation wage. This is because employees can quit.[10] It is assumed, though, that the firm cannot leave costlessly if a bad state of nature occurs, and hence requires only the average utility it would expect to get by switching into another sector.

What will the optimal employment contract be like? One possible form would be for the firm to pay a fixed wage across all states of nature, but a risk-averse firm is not going to be happy with such an agreement. It would prefer to share some of its risk. There is a way to achieve that, namely, by having a labor contract in which pay, w, is linked implicitly or explicitly to the number of jobs, n. Employment can be observed by both sides, so a contract of form $w(n)$ is feasible.

If the wage is a function of employment, profits can be written

$$\pi=\theta f(n)-w(n)n. \tag{1}$$

The firm's utility is $v(\pi)$, and it will be useful to define a function

$$r(n,\theta)=v(\theta f(n)-w(n)n), \tag{2}$$

which the firm is to maximize, by its choice of employment, once the product price θ is known. For well-behaved optima, therefore, the number of jobs are chosen so that

$$r_n=v'(\pi)[\theta f'(n)-w'(n)n-w(n)]=0 \tag{3}$$

and $r_{nn}<0$. Let $n(\theta)$ be the function which solves this. Then it turns out that employ-

[9]This is the same as the normal expected utility function (it is equivalent here to utilitarianism) where a worker has probability $1-n$ of being unemployed.

[10]The alternative assumption of immobility can be introduced simply by leaving out the constraint, and the main point of the paper is unchanged.

ment is an increasing function of the price of output, namely,

$$(4) \qquad n'(\theta) = -v'(\pi)f'(n)/r_{nn} > 0,$$

which will be used later on. Finally, because the optimization problem to come can be set up most concisely using a maximum value function, let

$$(5) \qquad \nu(\theta) = \max_{n} r(n,\theta).$$

By an envelope result, therefore,

$$(6) \qquad \nu'(\theta) = v'(\pi)f(n) > 0,$$

which measures the rise in the firm's *ex post* utility that is caused by an increase in the product price. The employer does better as states of nature improve.

The optimal form of employment contract is to be thought of, as usual, as the solution to a maximization problem. The group of workers—they might be taken to be a trade union, but that is not essential—must pick a wage function $w(n)$ which maximizes their expected utility subject to various constraints. This contract will not be first-best Pareto optimal, because of the existence of asymmetric information about conditions in the product market, but will in the obvious sense be a second-best optimum. There are various ways to characterize it mathematically. The simplest is to work not with a wage function defined on employment, but rather with two functions defined on the state of nature θ. Hence one could derive a function for pay, $w(\theta)$, and a function for jobs, $n(\theta)$, and let these together define implicitly the wage as a function of the employment level.[11] The paper uses a variant of this in which the choice variables are $\nu(\theta)$ and $n(\theta)$, but this is only a matter of convenience.

The full problem can be written as

$$(7) \qquad \max_{\nu(\theta),\, n(\theta)} \int_{\underline{\theta}}^{\bar{\theta}} \{nu(w) + (1-n)u(b)\}g(\theta)\,d\theta$$

subject to

$$(8) \qquad \int_{\underline{\theta}}^{\bar{\theta}} \nu(\theta)g(\theta)\,d\theta \geq \nu^*$$

$$(9) \qquad \nu'(\theta) = v'(\pi)f(n)$$

$$(10) \qquad u(w) \geq u(b),$$

where for brevity the wage, employment, and profit variables have not been written explicitly as functions of θ. The maximand is the expected utility of the labor group, and the choice variables here are the utility of the firm and the number of jobs. Equation (8), the first of the three constraints, states that the employer's expected utility (when setting employment optimally) must be at least what it can achieve in another sector of the economy. The second constraint, equation (9), captures the fact that the firm will move on to its labor demand curve once the price of output is known. It implies and is implied by the condition that the value of the marginal product of labor be equal to its marginal cost. Equation (10), the third of the constraints, stipulates that an individual worker be no worse off, *ex post*, than he or she would be by drawing government unemployment benefit. This is a feasibility condition.

The firm's utility from profits, $v(\pi)$, is monotonic, so can be inverted. At the optimum the employer's utility is $\nu(\theta)$. Thus it is possible to define profits by the function $\pi(\nu(\theta))$, which is the inverse of $v(\pi)$. The product of the derivatives of these two functions is unity. The relationship $\pi(\nu)$ provides a way to write the wage as a function of the realized product price. It is, by rearrangement of the definition of profits,

$$(11) \qquad w = (\theta f(n) - \pi)/n = (\theta f(n(\theta)) - \pi(\nu(\theta)))/n(\theta).$$

These points mean that both the wage and the profit level can be substituted out of the optimization problem described by equations (7) to (10).

The problem will be solved here by integrating equation (9) by parts and using the result to form a Lagrangean. Let the multiplier on the first constraint be λ, that on the

[11] See Green and Kahn, for example.

second be $\psi(\theta)$, and that on the third be $\phi(\theta)$. The Lagrangean is then

$$(12)\quad L=\int_{\underline{\theta}}^{\bar{\theta}}\{[nu(w)+(1-n)u(b)+\lambda\nu(\theta)]\}g(\theta)-\nu(\theta)\psi'(\theta)-\psi(\theta)v'(\pi)f(n)+\phi(\theta)[u(w)-u(b)]\}\,d\theta.$$

There are, strictly speaking, some endpoint terms from the integration by parts, but for simplicity they have been omitted. The first-order conditions for an interior optimum include[12]

$$(13)\quad \nu(\theta):\ [\lambda-(u'(w)/v'(\pi))]g(\theta)-\psi'(\theta)-(v''(\pi)/v'(\pi))\psi(\theta)f(n)-\phi(\theta)u'(w)/(v'(\pi)n)=0$$

$$(14)\quad n(\theta):\ \{u(w)-u(b)+[\theta f'(n)-w]u'(w)\}g(\theta)-\psi(\theta)v'(\pi)f'(n)+(\phi(\theta)u'(w)/n)[\theta f'(n)-w]=0$$

$$(15)\quad \psi(\underline{\theta})=\psi(\bar{\theta})=0$$

$$(16)\quad u(w)-u(b)\geq 0 \qquad \phi(\theta)\geq 0$$

Equation (15) gives the transversality conditions: the multiplier $\psi(\theta)$ must start and end at zero. Equation (16) is the complementary slackness condition on (10): the multiplier $\phi(\theta)$ is zero when the inequality $u(w)\geq u(b)$ fails to bind.

III. Some Results

We can now derive some of the model's properties. The analysis will concentrate on the case in which constraint (10) is not strictly binding (it is straightforward to incorporate).

The first points that need to be established are about the sign and behavior of the multiplier $\psi(\theta)$ which corresponds to the differential equation (9). It turns out to be possible to show that $\psi(\theta)$ cannot be negative. It must begin at zero, become positive for some states of nature, and then return to zero for the best state of nature. The proof (given in the Appendix) shows that once $\psi(\theta)$ becomes negative it remains negative. This would violate the transversality condition $\psi(\bar{\theta})=0$, which is what makes it possible to rule out negative values for the multiplier.

The model can generate equilibria in which there is both involuntary unemployment and underemployment. Involuntary unemployment will be defined here as an outcome in which the utility from work, $u(w)$, is strictly greater than the utility from being unemployed, $u(b)$. This seems to be the most natural definition. It captures the idea that there can be an individual without a job who would like to change places with a similar individual who is employed. A definition of underemployment and overemployment is also needed, and the normal one will be used. The value of the marginal product of labor is $\theta f'(n)$; the value of the reservation wage is b. Overemployment will be said to exist when $\theta f'(n)<b$ and underemployment will be said to exist when the inequality is reversed. The intermediate case is, as we should expect, the efficient competitive outcome $\theta f'(n)=b$. Conventional implicit contract theory solves the problem

$$(17)\quad \max_{w(\theta),\,n(\theta)}\int_{\underline{\theta}}^{\bar{\theta}}[nu(w)+(1-n)u(b)]g(\theta)\,d\theta$$

subject to

$$(18)\quad \int_{\underline{\theta}}^{\bar{\theta}}[\theta f(n)-wn]g(\theta)\,d\theta\geq\pi^*,$$

with first-order condition

$$(19)\quad u(w)-u(b)-u'(w)w=-u'(w)\theta f'(n).$$

[12] These presuppose that the problem is well-behaved.

By adding $bu'(w)$ to both sides we have

$$(20)\quad [u(w)-u(b)]+u'(w)(b-w) = u'(w)[b-\theta f'(n)] \geq 0,$$

which is positive by concavity of the $u(\cdot)$ function. The production function $f(n)$ is also concave, so that $b \geq \theta f'(n)$ implies that employment exceeds that at the point at which $b=\theta f'(n)$. This is the overemployment result of early contract theory. Notice, however, that involuntary unemployment ($u(w)>u(b)$) can certainly exist. The second generation of work on labor contracts (see Grossman and Hart, for example) used an amended version of the framework developed earlier. In this case private unemployment insurance is assumed to exist and to increase b to the point at which there is full insurance. Underemployment, $\theta f'(n)-b>0$, does occur, but it is all voluntary.

It is not surprising that, by taking an element from each of the two generations of contract theory, we can produce a model in which there are insufficient jobs and dissatisfied unemployed workers. The proof uses equation (14). It can be rewritten, by adding $g(\theta)bu'(w)$ to both sides, and imposing the assumption $\phi=0$, as

$$(21)\quad g(\theta)[u(w)-u(b)+u'(w)(b-w)] = u'(w)g(\theta)[b-\theta f'(n)] + \psi(\theta)v'(\pi)f'(n).$$

Concavity of the worker's utility function again ensures that the left-hand side is positive. In this model, however, it is not the case that overemployment follows immediately. The reason is that $\psi(\theta)$ is necessarily positive, so that underemployment can exist at the same time as equation (21) is satisfied. It is the mixture of the assumptions of asymmetric information and imperfect unemployment insurance which generates the possibility that underemployment and involuntary unemployment can occur simultaneously.

Some other results can be established. First, there exist at least two intervals, $[\underline{\theta},\theta^1]$ and $[\theta^2,\bar{\theta}]$, on which there is overemployment. Equation (21) makes it clear that when $\psi(\theta)$ is zero, the usual overemployment conclusion (as in equation (20)) goes through. That must occur, by the transversality condition, in the best and worst states of nature. By continuity, therefore, there exist intervals at either end of the support of $g(\theta)$ on which the value of the marginal product of labor drops below the reservation wage. Second, the same argument can be used to prove that, for the lowest price ($\underline{\theta}$) and the highest price ($\bar{\theta}$), the model simply replicates the equilibrium of a Baily-Azariadis-Gordon model. This is formally equivalent to the result in optimal non-linear tax theory (James Mirrlees, 1976, for example) that the marginal tax rate should be zero at the top and bottom ends of the skill distribution. Third, equation (19) reveals that, as is well known, in first-generation contract theory, the wage exceeds the value of the marginal product of labor. Yet this is no longer true once asymmetric information is introduced. Inspection of equation (21) shows that $w>b$ and $w\leq \theta f'(n)$ can hold simultaneously.

It is possible to provide a more intuitive account of the forces which lead to overemployment or underemployment. The first is a response to a missing insurance market. By assumption, in this and early contract models, there is no private unemployment insurance scheme to allow workers to avoid the risk of a drop in utility if fired. Workers must therefore find some other way, if they can, to guard against the vagaries of the market. The optimal one is to ensure that the firm takes on more employees than would be justified on the grounds of technical efficiency. That reduces the risk of unemployment. The greater is workers' risk aversion, the larger will be the extent of the overemployment. Underemployment has quite a different cause. When there is asymmetric information about prices, the optimal labor contract ties pay to the number of jobs, which is observable and carries some information about the employer's success in the product market. The wage is an increasing function of employment whenever, at the optimum, the value of the marginal product of labor exceeds the wage rate. Equation (3)

states this formally. In this case, however, there is a reason to employ fewer individuals than would be optimal in the equivalent market with a competitively fixed wage: the marginal cost of labor is higher than in the equivalent market with wage taking by firms. This apparent inefficiency (it is second-best efficient) is the indirect result of the fact that workers do not have access to the same information as their employer.

IV. Why Is There Not More Private Unemployment Insurance?

The data given earlier in the paper suggest that even in union contracts there is often no private unemployment insurance (SUB) scheme. It is natural to try to explain why. The model given in the paper does not do so; it relies on an appeal to the empirical evidence.

There are at least five broad possibilities. First, layoffs are not random, as this paper and most of the literature have assumed, but are typically made according to the rule that juniors be dismissed first. In a world with "layoffs by inverse seniority," there may be good reason to think that private unemployment insurance will be uncommon. It might be argued that the typical senior employee, whose job is safe in all but the most exceptional recessions, has little incentive to make contributions into a fund which will pay money to those juniors who lose their jobs.[13] Second, it may be that the public provision of unemployment insurance has largely eliminated private schemes. This seems especially likely in cases in which public payments are compulsorily reduced if the recipient has income from a private fund. Third, if insurers have very little information about the probability distribution of unemployment states, and workers are not particularly risk averse, we might expect the insurance market to be thin. Fourth, traditional moral hazard arguments apply. It is likely to be difficult for an insurance company to check the motives of those who enter the unemployed pool.[14] Fifth, if product market shocks are described by an autoregressive process, and capital markets are less than perfect, an insurance company would often face "gambler's ruin."

This paper pursues only the first of these ideas. That is because it fits in with the implicit methodological approach taken earlier (that contract theory should take more account of what real labor contracts are like), and because the mathematical structure is similar to that already used.

A. *A Seniority Model*

Most of the previously made assumptions will apply again. It is the form of the working group's objective function which must be altered. The most natural course is to assume majority voting and to let the new maximand be the expected utility of the typical senior worker. The mathematics is much simplified if we concentrate on states of nature in which less than half the work force is laid off. This seems only moderately restrictive,[15] and produces a great gain in tractability. The assumption guarantees that there is a majority of workers, all of whom have the same utility function, whose preferences can be taken to be those of the labor group. Under layoffs by seniority, therefore, the optimal employment contract is the solution to the following problem:

$$\max \int u(w) g(\theta)\, d\theta \tag{22}$$

subject to

$$\int v(\theta) g(\theta)\, d\theta \geq v^* \tag{23}$$

$$v'(\theta) = v'(\pi) f(n) \tag{24}$$

$$u(w) - u(b) \geq 0, \tag{25}$$

[13] One can think of some objections to this. First, in massively declining industries even seniors' jobs are not safe. Second, seniors may be altruistic. Third, if workers earn no rents, it may be optimal—to ensure that there is a supply of new recruits—for senior employees to subsidize young workers who are without work.

[14] Richard Arnott and Joseph Stiglitz (1982) discuss these issues.

[15] The first-order conditions derived here will continue to hold, over these states of nature, in a more general model. Without state-contingent contracts, however, things would be quite different, and this avenue seems worth pursuing.

where the maximand is to be thought of as the utility of a senior worker, who cares only about the wage rate (by virtue of the insulation from layoffs that is produced by the seniority ordering). The choice variables are again $\nu(\theta)$ and $n(\theta)$.

Write the Lagrangean as

$$(26)\quad L=\int\{[u(w)+\lambda\nu(\theta)]g(\theta)-\psi'(\theta)\nu(\theta)+\phi(\theta)u(w)-\psi(\theta)v'(\pi)f(n)\}\,d\theta.$$

The new first-order conditions are

$$(27)\quad \nu(\theta):\ \left[\lambda-\frac{u'(w)}{nv'(\pi)}\right]g(\theta)-\psi'(\theta)-\frac{u'(w)\phi(\theta)}{nv'(\pi)}-\psi(\theta)\frac{v''(\pi)}{v'(\pi)}f(n)=0$$

$$(28)\quad n(\theta):\ (1/n)u'(w)[g(\theta)+\phi(\theta)]\times[\theta f'(n)-w]-\psi(\theta)v'(\pi)f'(n)=0.$$

The transversality and complementary slackness conditions (equations (15) and (16)) still apply. It is no longer possible to sign unambiguously the multiplier $\psi(\theta)$, although it is to be expected that $\psi(\theta)$ would be positive over much of the interval $[\underline{\theta},\bar{\theta}]$. This means that there are fewer definite conclusions than before.

Perhaps the most striking feature of this case is a characteristic of equilibria near the ends of the probability distribution. As long as the feasibility condition $u(w)-u(b)\geq 0$ is not strictly binding, $\phi(\theta)$ disappears. Because $\psi(\underline{\theta})=\psi(\bar{\theta})=0$, equation (28) then simplifies to

$$(29)\quad \underline{\theta}f'(n(\underline{\theta}))-w(\underline{\theta})=\bar{\theta}f'(n(\bar{\theta}))-w(\bar{\theta})=0.$$

Hence the wage is equal to the value of the marginal product of labor in the worst and best states of nature. More generally, involuntary unemployment (among juniors) can still occur, and employment may be above or below the level at which the value of the marginal product of labor equals the reservation wage.

V. Conclusion

The strengths of traditional implicit contract theory (Baily; Gordon; Azariadis, 1975) were that it could explain, first, wage ridigity and, second, involuntary unemployment (in the sense that unemployed workers receive less utility than identical employed workers). This has been supplanted, in the early 1980's, by the theory of labor contracts under asymmetric information. The strengths of this literature are that it is more general and that, unlike early contract theory, it can explain employment levels below that at the competitive equilibrium. We now know that wage rigidity is, in either model, a special case which is generated by risk neutrality of employers,[16] or by even more restrictive assumptions. On the issue of unemployment, however, the two approaches differ. Early theory can generate equilibria with involuntary unemployment, but at employment levels greater than in perfect spot markets. Later theory, like the many-worker model of Grossman and Hart, can produce equilibria with inefficiently high unemployment, but it is all voluntary unemployment.

This paper has tried to combine the desirable features of both approaches. The model which results has different properties from those in the literature, and can lead, in particular, to equilibria in which there is both involuntary unemployment and inefficiently high unemployment. One assumption is central to this: it is that workers do not receive unemployment insurance from their employers. Azariadis, Baily, and Gordon originally made this assumption, and so does Holmstrom. Grossman and Hart, and Hart, among others, make the opposite assumption, that there is a perfect layoff insurance scheme.

[16] Baily and Azariadis both assume that layoffs are by random draw, and the proof of the famous wage rigidity theorem relies on that. Yet we know this assumption to be unrealistic, and the obvious alternative of layoffs by seniority generates models (see my earlier study, 1984) in which the wage rigidity result does not hold.

It is not easy to know whether models of labor contracts under asymmetric information should or should not assume that workers receive private unemployment insurance. As I have shown, however, the results are very sensitive to this assumption, which is one reason to examine its implications. There is another reason. In Section I, I presented some apparently little known data on the extent of unemployment insurance schemes in U.S. labor contracts. The figures, covering roughly six million workers, reveal that half of the employees in manufacturing industry were covered by a Supplemental Unemployment Benefit plan. This is probably higher than most economists realize, but it still leaves a huge number of individuals without insurance coverage. In nonmanufacturing industry the figures were starker: SUB plans hardly existed. Severance pay and wage-employment guarantees were more common than SUB arrangements, but still only affected part of the nonmanufacturing labor force. Although one can only guess, these figures suggest that most unionized workers in the United States do not, in fact, receive the kinds of unemployment insurance payments assumed in recent work on the theory of labor contracts. Nonunion workers are probably even less well insured. Hence there may be some point to models of the kind set out in this paper.

Appendix

A Proof of the Nonnegativity of $\psi(\theta)$

The analysis of the main section of the paper relies on the fact that the multiplier $\psi(\theta)$ is weakly positive. The proof is as follows.

Equation (13) is a linear differential equation of first order. It can be rewritten—setting ϕ to zero on the assumption that (10) does not bind—as

$$\text{(A1)}\quad \psi'(\theta)=\left[\lambda-\frac{u'(w)}{v'(\pi)}\right]g(\theta)+\psi(\theta)\rho f(n),$$

where $\rho\equiv-v''(\pi)/v'(\pi)$. The solution to the equation can be found by integrating it using the condition $\psi(\underline{\theta})=0$. That solution is

$$\text{(A2)}\quad \psi(\theta)=\int_{\underline{\theta}}^{\theta}\left[\lambda-\frac{u'(w)}{v'(\pi)}\right]g(x)H(x)\,dx$$

where

$$\text{(A3)}\qquad H\equiv\exp\int_{\theta}^{x}-\rho f(n)\,dz,$$

and x and z are used to denote the variable of integration.

The endpoint conditions, equation (15), also ensure that $\psi(\bar{\theta})=0$, which is the key to the proof of the multiplier's nonnegativity. Define a new variable

$$\text{(A4)}\qquad A(\theta)\equiv\lambda-u'(w)/v'(\pi).$$

If $\psi(\theta)$ is to satisfy the transversality requirement, $A(\theta)$ must change sign over the interval $[\underline{\theta},\bar{\theta}]$. The derivative of that function is

$$\text{(A5)}\quad A'(\theta)=-\frac{u''(w)}{v'(\pi)}\frac{dw}{d\theta}+\frac{u'(w)v''(\pi)}{[v'(\pi)]^2}\frac{d\pi}{d\theta}.$$

By earlier definitions, $w=(\theta f(n)-\pi)/n$, $\pi=\pi(\nu)$ and $\nu'(\theta)=v'(\pi)f(n)$. Hence

$$\text{(A6)}\quad A'(\theta)=-\frac{u''(w)}{nv'(\pi)}[\theta f'(n)-w]n'(\theta)-\rho(u'(w)/v'(\pi))f(n).$$

The derivative $n'(\theta)$ is known to be positive. Hence negativity of $A'(\theta)$ would be guaranteed by the condition $w>\theta f'(n)$.

I wish to prove $\psi(\theta)\geq 0$. Assume not, in order to obtain a contradiction. Equation (14) in the text then implies that $[\theta f'(n)-w]$ is negative. This ensures $A'(\theta)<0$. Thus, once $\psi(\theta)$ takes a negative value, it remains negative. For an optimum, however, $\psi(\theta)$ must return to zero at $\theta=\bar{\theta}$. Hence $\psi(\theta)<0$ implies a contradiction.

REFERENCES

Akerlof, George A. and Miyazaki, Hajime, "The Implicit Contract Theory of Unemployment Meets the Wage Bill Argument," *Review of Economic Studies*, March 1980, *47*, 321–38.

Arnott, Richard and Stiglitz, Joseph E., "Equilibrium in Competitive Insurance Markets: The Welfare Economics of Moral Hazard," mimeo., Queen's University, 1982.

Azariadis, Costas, "Implicit Contracts and Underemployment Equilibria," *Journal of Political Economy*, December 1975, *83*, 1183–1202.

______, "Employment with Asymmetric Information," *Quarterly Journal of Economics*, Supplement 1983, *98*, 157–72.

Baily, Martin N., "Wages and Employment Under Uncertain Demand," *Review of Economic Studies*, January 1974, *41*, 37–50.

Beal, E. F., Wikersham, E. D. and Kienast, P., *The Practice of Collective Bargaining*, Chicago: Richard D. Irwin, 1972.

Calvo, Guillermo and Phelps, Edmund, "Employment Contingent Wage Contracts," *Journal of Monetary Economics*, Supplement 1977, *3*, 160–68.

Gordon, Donald F., "A Neoclassical Theory of Keynesian Unemployment," *Economic Inquiry*, December 1974, *12*, 431–59.

Green, Jerry and Kahn, Charles, "Wage-Employment Contracts," *Quarterly Journal of Economics*, Supplement 1983, *98*, 173–82.

Grossman, Gene M., "Union Wages, Temporary Layoffs, and Seniority," *American Economic Review*, June 1983, *73*, 277–90.

Grossman, Sanford J. and Hart, Oliver D., "Implicit Contracts, Moral Hazard and Unemployment," *American Economic Review Proceedings*, May 1981, *71*, 301–07.

Hall, Robert E. and Lilien, David M., "Efficient Wage Bargains under Uncertain Supply and Demand," *American Economic Review*, December 1979, *69*, 868–79.

Hart, Oliver, D., "Optimal Labour Contracts under Asymmetric Information: An Introduction," *Review of Economic Studies*, January 1983, *50*, 3–36.

Holmstrom, Bengt, "Equilibrium Long-Term Labor Contracts," *Quarterly Journal of Economics*, Supplement 1983, *98*, 23–54.

Mirrlees, James A., "Optimal Tax Theory: A Synthesis," *Journal of Public Economics*, November 1976, *6*, 327–58.

Oswald, Andrew J., "Efficient Contracts are on the Labor Demand Curve: Theory and Facts," Working Paper No. 178, Industrial Relations Section, Princeton University, 1984.

U.S. Bureau of Labor Statistics, *Characteristics of Major Collective Bargaining Agreements*, Washington, various issues.

The Design of Procurement Contracts

By RAFAEL ROB*

Defense contracting has traditionally been and is expected to remain a major component of the U.S. government budget. During 1983 the Department of Defense (DOD) spent a grand total of $53.6 billion on procurement programs which amounted to 24.6 percent of the defense budget (see *Budget of the U.S. Government, FY 1985*). Projected figures for future years are even higher. For the 1987 fiscal year it is estimated that $106.7 billion will be appropriated for the same purpose. It will represent 31.2 percent of the budget. By comparison, the corresponding figures for the fiscal year of 1976 were $16 billion (19.7 percent).

Given these numbers, the importance of cost effectiveness in the acquisition process cannot be overemphasized. Clearly the public shares this view and has repeatedly expressed it through the popular press and other media channels. A case in point is the purchase of spare parts and services for major defense systems. It is often alleged that these goods and services are grossly overpriced, that their exaggerated cost is sustained by a sole-source contracting arrangement, and that competitive practices could generate considerable price reductions. These claims are customarily substantiated by exhibiting competing firms capable of producing virtually the same items for a small fraction of what the taxpayer is currently paying for them. A popular counterargument to such an allegation is that a costly research and development ($R\&D$) stage precedes the construction phase of defense systems. The contractor subsequently imputes these preliminary expenses into the price; and the price he ends up charging reflects historical (*ex post*) costs. This compensation provides the necessary inducement for undertaking the correct amount of $R\&D$ activity.

The objective of this paper is to examine this issue. I argue that savings can be realized if only a fraction of a project is awarded to a single producer while the remainder is competitively purchased upon termination of the original contract. Such a procedure is known as an "educational" or "learning" buy. Indivisibilities are frequently of no consequence here since the DOD normally contemplates the purchase of many units. But even if lumpiness is an inherent problem, such as when elaborate weapon systems are ordered, it is still possible to purchase replacement parts from other suppliers. This will have the same cost-reduction effect as splitting up a project.

The effect of an educational buy is to reinforce competition by alleviating the initial cost uncertainty which keeps many small firms out of the defense industry. Once the $R\&D$ stage is completed, costs and technical designs are known and new firms enter into the industry. This enables the government to purchase the remaining units (or subsystems) at relatively low prices. Surely, a commitment to buy smaller quantities will lessen the incentives of the original contractor to invest in $R\&D$ at the first stage; but when competition is lacking (and this is certainly the case in the defense industry), the price of incremental $R\&D$ is exceedingly high while the creation of more competition is, as I shall show, a superior alternative.

The above recommendation is derived in a model where firms submit sealed bids for a so-called firm fixed price (*FFP*) contract. That is, potential producers promise to deliver the systems for a predetermined per unit price, no matter what their actual expenses turn out to be. The assumption underlying the analysis is that firms are risk neutral. Under this assumption, the *FFP*

*Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104. An earlier version of this paper was delivered at the April 1985 INRIA Conference in Paris. The research was partially funded by the National Science Foundation. Helpful comments from two anonymous referees and John Riley are gratefully acknowledged.

arrangement is an optimal incentive contract. Of course, when producers are risk averse, the problem of tradeoff between risk-bearing and cost-reduction incentives emerges. This problem has been dealt with in Martin Weitzman (1980). Therefore, I will proceed under the risk-neutrality assumption.

The rationale behind this structure is to bring together two major issues that characterize the weapon acquisition process. The first is the problem of choosing a most efficient producer among a group of heterogeneous contenders differentiated by their cost effectiveness. Formally, this issue is akin to optimal auctions theory (see John Riley and William Samuelson, 1981). Accordingly, the latter constitutes an important building block of my formulation. The second issue is the problem of inducing the chosen contractor to pursue a socially desirable research and development strategy. This is essentially an agency-type problem (see the papers by Michael Cummins, 1977; Charles Holt, 1979; F. M. Scherer, 1964; and Weitzman). These two problems were hitherto treated separately and the model below is an attempt to remedy this shortcoming.

A second feature of the model below is that it is explicit about the *R&D* process. The firm attempts to reduce its production costs by experimenting with (or sampling from) new techniques. Sampling is a costly activity, the outcomes of which are unforeseen. This approach is intended to capture the element of cost uncertainty to which many practitioners in this area often allude. The important contribution of Holt should be recalled in this regard. He satisfactorily addressed the issue of auctioning governmental contracts under conditions of cost uncertainty. The element of uncertainty on which his study focused, however, is purely exogenous (see condition 3 of his Proposition 1). In particular, he assumed away the choice of cost-reducing efforts by the winning firm and the associated moral hazard problem. As mentioned, I shall consider the auctioning *and* the incentive problems jointly.

Third, the notion that an ongoing relationship where the firm delivers goods and services over time is integrated into the structure. In fact, the implementation of a learning buy is largely facilitated by the existence of a preliminary *R&D* stage, a time span of delivery, and the possibility of separating the purchase of a system from its individual components.

I. The Model

I first derive the research strategy undertaken by a firm which has already won an open public auction to produce y units of a given defense system. As a result of a preliminary bidding procedure (which is later specified), it committed itself to deliver the product for a fixed per unit price, p. Before construction starts, production costs, c, are determined by experimenting with new techniques (the *R&D* stage). It is assumed that experimentation is like sampling with perfect recall from a known distribution with a cumulative distribution function given by $G(c)$. That is, if the firm decides to stop experimentation after n samples have been taken, its unit cost will be those of the lowest cost technique it is aware of at this stage. Its profits will accordingly be

$$(1)\qquad \pi = y(p-c_m)-ns,$$

where $c_m \equiv \mathrm{Min}\{c_1,\dots,c_n\}$ and s stands for its firm-specific sampling (or learning) cost.

Search theory (see John McCall, 1970a, b) tells us that the contractor will optimally experiment (or search) until the expected gain from searching *one* more time drops below the cost of search, s. Formally, using the dynamic programming formulation, note that (n, c_m) are the relevant state variables and let $v(n, c_m)$ denote the value function of the firm. This v function must satisfy

$$(2)\qquad v(n,c_m) = \mathrm{Max}\{y(p-c_m)-ns,\; E_c v(c_m \Lambda c, n+1)\}.$$

where $a\Lambda b \equiv \mathrm{Min}\{a,b\}$. It is not hard to confirm that the solution to the functional equation (2) is given by

$$(3)\qquad v(n,c_m)=\begin{cases} y(p-c_m)-ns, & \text{if } c_m<z\\ y(p-z)-ns, & \text{if } c_m>z\end{cases}$$

where z is defined by

$$(4a)\quad y\tilde{G}(z)=s$$

$$(4b)\quad \tilde{G}(z)=\int_0^z G(c)\,dc.$$

Intuitively, z is that cost level where the contractor is indifferent between stopping and searching once more. As noted, the optimal strategy for the firm is to keep experimenting until its per unit production costs are reduced below the cutoff level z. Note that z depends on the datum of the problem (y,s) in a rather intuitive way. Specifically, z is increasing in s, indicating that the less efficient a firm is (the higher is its s), the earlier on will it stop its research endeavors. Similarly, z is decreasing in y; that is, a larger contract induces greater incentives to engage in *R&D*. Formally, we obtain by implicit differentiation:

$$(5a)\quad \partial z/\partial s=1/yG(z),$$

$$(5b)\quad \partial z/\partial y=-\tilde{G}(z)/yG(z).$$

For future reference, let us denote the *average actual* cost associated with the cutoff level z by $a(z)$. By definition:

$$(6)\quad a(z;y,s)=\int_0^z c\,dG(c)/G(z)$$
$$=z-\tilde{G}(z)/G(z)=z-s/yG(z),$$

where the last equality follows from (4).

Given the above research strategy, we next consider the firm's profits before the *R&D* stage. *Ex ante*, that is, prior to knowing what realized costs are going to be, we have

$$(7a)\quad \pi_0=E_c v(1,c)=Pr(C_1>z)(\pi_0-s)$$
$$+y\int_0^z(p-c-s)\,dG(c)$$
$$=[1-G(z)]\pi_0+y\int_0^z(p-c)\,dG(c)-s$$
$$=[1-G(z)]\pi_0+y(p-z)G(z)$$
$$+\int_0^z G(c)\,dc-s.$$

Thus

$$(7b)\quad \pi_0=y(p-z).$$

Assume now that even the most efficient firm (i.e., the one with the lowest s) encounters incipient competitors who can produce almost as efficiently. Some of them can always undercut it when it makes above normal profits. In such an industry, profits are driven to zero and the market price of a procurement contract, p, is

$$(8)\quad p=z.$$

The right-hand side of this equation is to be understood as the z value corresponding to the most efficient ($\equiv$ lowest s) producer.

Having said this about firms' behavior and the resulting equilibrium price, let us focus our attention on what the government can do to minimize its outlays on the project in question. While for individual producers the size of the project, y, is parametrically given, for the government it is a choice object. The government, in other words, can initially enter into a contractual arrangement with a single producer. This contractor promises to deliver a prespecified quantity, y, for a given price, p. He also agrees to disclose his production-induced technological knowledge so that any firm can later adapt the technology that results from his *R&D* efforts.

Clearly, these other firms are prone to be less efficient than the original contractor because of the creation of firm-specific knowledge and because technological information, experience, and learning are only fractionally transferable between firms. One obvious facet of this "interface problem" is that the original contractor will try to keep his competitive edge by hiding any information which may be useful for the development of new marketable products (see Oliver Williamson, 1967). For modeling purposes, I shall assume that there is an interface efficiency loss of $(100(\alpha-1))$ percent (with $\alpha>1$). That is, that the unit costs of other producers are αc_m (rather than c_m). This interface problem is the primary reason for using a single contractor for both the *R&D* and some of the production.

After the educational buy is consummated, firms can bid for any future contracts. At this point, advertised competition is implemented and prices are driven to their runner-up cost, αc_m. Note that at the bidding stage, c_m is unknown. Its *ex ante* expected value is what I denoted by $a(z(y,s); y,s)$ (see equation (6)).

When only a fraction of a project is acquired from one contractor (or, for that matter, when spare parts and services are bought from other sources), a certain tradeoff is involved. Clearly, the smaller the quantity the government is committed to buy from a given producer, the less are his incentives to lower the cost of the product. This is so because when the contractor cannot fully appropriate the value of his *R&D* expenditures, his investments are going to be smaller (see equation (5b)). On the other hand, by eliminating the contractor as a sole source producer, competition is enhanced and prices decrease. Our goal here is to see whether and under what circumstances the government can exploit this tradeoff.

Based on these considerations, we may write the governmental objective as

$$TC \equiv yp + \alpha(1-y)a(z). \tag{9}$$

The form of this criterion incorporates the assumption that the overall size of the project is fixed in advance. Setting this size equal to one is just a convenient normalization. This assumption is clearly restrictive since the outcomes of research and development would in themselves dictate how many units the DOD wants to purchase.[1] Specifically, lower realized production costs (c_m) would induce larger orders. In Section IV below, I shall extend the analysis to cover this case.

II. Analysis

A. *Large Number of Bidders*

Under perfect competition, that is, when (8) is satisfied, the objective, (9), becomes

$$TC = yz + \alpha(1-y)a(z). \tag{9$'$}$$

[1] I thank an anonymous referee for pointing out this fact.

Differentiating (9′) we obtain

$$\begin{aligned} z + y(\partial z/\partial y) - \alpha a(z) \\ + \alpha(1-y)a'(z)(\partial z/\partial y) \\ = -(\alpha-1)[z-(\tilde{G}(z)/G(z))] \\ - \alpha((1-y)/y)a'(z)(\tilde{G}(z)/G(z)), \end{aligned} \tag{10}$$

where (5) and (6) have been used to derive the last equality. Looking at (10) we observe that the objective TC diminishes throughout the interval (0,1). It follows that a corner solution at $y=1$ is a global minimum, that is, that the government can do no better than award the full project to a single firm. Thus we have

PROPOSITION 1: *If perfect competition prevails both at the initial contracting stage and at the production stage, then governmental costs are minimized when a single producer is contracted to deliver the entire system.*

B. *Small Number of Bidders*

The result above depends critically on the assumptions I have made along the analysis. Most importantly, it hinges on the zero profits assumption. It is well known that for any given project the number of potential contractors very rarely exceeds five (see Larry Yuseph, 1976). Moreover, each producer knows not just the technological facts and the contingencies it might face in developing a defense system, but also the minimal price at which he is willing to accept the contract. Given this privacy of information and the limited number of competitors, producers are expected to bid for positive profits. That is, we expect p to be greater than z so that equation (8) is violated. Motivated by these considerations, let us reanalyze problem (9) under conditions of imperfect competition.

To do that, envision a finite number of firms, n, bidding for the contract. The DOD will accept the lowest bid and the items are to be delivered at this price. Competing producers are differentiated by their *R&D* cost conditions, that is, by their s parameter and

thus by their implied z value. We now think of z as a firm-specific reservation value and write $z_1, \ldots, z_n$. By (7), z_i is the lowest price at which the ith producer is willing to undertake the project. In line with my previous discussion, it is assumed that the exact value of s is private information and thus that the reservation values, z_i's, are private information as well. Denote by $H(\cdot)$, $h(\cdot)$ the cumulative distribution function and probability density function, respectively, of s. The variable H reflects the probabilistic beliefs (held by the government and competing firms) about others' cost conditions. It embodies in it everything that is common knowledge about the $R\&D$ process.

The cutoff rule, (4), together with H induces a distribution over the reservation values, z_i. Notationally, call F the complementary (when we cumulate from the right) cumulative distribution function of Z, that is, define

$$F(z) = Pr(Z > z) = 1 - Pr(Z \leq z).$$

The induced F is then

$$(11) \qquad F(z; y) = 1 - H(y\tilde{G}(z)).$$

We can now forget about s and conduct the analysis in terms of the z parameter instead. Accordingly, I shall work directly with the distribution of these characteristics, F.

Next, the design of a cost-minimizing bidding procedure is formalized. In doing so, I closely follow the approach taken by Riley and Samuelson. My presentation is thus brief and many of the details are relegated to the Appendix. An example illustrating the succeeding analysis is provided at the end.

Let $b(z)$ $(= b(z; y))$ be the common symmetric bidding strategy of the n producers when a project of size y is initially contracted out. If all but one firm adopt $b(\cdot)$, it would clearly be in the best interest of the remaining firm (the first one, say) to bid somewhere in the range of $b(\cdot)$. Thus, its bid, b_1, can be expressed as $b_1 = b(x)$, for some x. Denote by $p(b_1, \ldots, b_n)$ the per unit price of the contract for firm 1. In a low-bid auction:

$$(12) \quad p(b_1, \ldots, b_n) = \begin{cases} b_1, \text{ if } b_1 < \text{Min}\{b_i\}_{i=2}^n \\ 0, \quad \text{otherwise.} \end{cases}$$

Let

$$\begin{aligned}(13) \quad P(x) &\equiv E_{z_2, \ldots, z_n} p(b(x), \\ &\qquad b(z_2), \ldots, b(z_n)) \\ &= E[b(x), x < \text{Min}\{z_i\}_{i=2}^n] \\ &= b(x) Pr(\text{winning the auction})\end{aligned}$$

be the *expected* price paid to a representative bidder. Using this notation, the expected gain to a type z_1 firm which bids as though it were a type x firm is (using (12) and (7))

$$\begin{aligned} &E[y(b(x) - z_1), \; x < \text{Min}\{z_i\}_{i=2}^n] \\ &= y\{E[b(x), \; x < \text{Min}\{z_i\}] \\ &\qquad - E[z_1, \; x < \text{Min}\{z_i\}]\} \\ &= y[P(x) - z_1 F^{n-1}(x)], \end{aligned}$$

where the last equality follows from (13). Since for the producer y is exogenously given, his optimal bidding problem may be posed as

$$(14) \quad \underset{x}{\text{Max}} \; \pi(x, z_1) = P(x) - z_1 F^{n-1}(x).$$

Since $b(\cdot)$ is the equilibrium bid strategy, agent 1's best reply must be $b_1 = b(z_1)$. In other words, the maximizing value x^* of (14) must be $x^* = z_1$. In the Appendix, I show that taking account of this constraint and of the cost of purchasing the remaining $(1 - y)$ units leads to the following governmental objective function:

$$\begin{aligned}(15) \quad TC(z_*, y) &= ny \int_{\underline{z}}^{z_*} F^{n-1}(z)[z\bar{f}(z) \\ &\qquad + \bar{F}(z) - \alpha a(z)\bar{f}(z)]\, dz \\ &+ n \int_{\underline{z}}^{z_*} \alpha a(z) F^{n-1}(z) \bar{f}(z)\, dz + z_0 F^n(z_*),\end{aligned}$$

where $\bar{F}=1-F$, $\bar{f}=-f$, $\underline{z}=$ lowest possible z and z_* is the maximum entry level. That is, the maximum reservation value at which it pays to enter into the auction. Firms whose minimum acceptance price exceeds z_* will drop out of the bidding game. In practice, the DOD can enforce z_* either by charging a bid submission fee which will deter certain firms from entering into the auction, or by announcing a maximum price at which contracts are going to be awarded. z_0 represents a governmental fall-back option. It corresponds to the next-best alternative for purchasing the system. It could, for instance, be the z value associated with "in-house" production or a previous contractor. Observe that an alternative derivation of (15) is through the "relevation principle" as in Roger Myerson (1981).

The form of the objective function shows that the problem of designing an auction by a procuring agent is reduced to a choice over two parameters only. One is the quantity ordered (y) from an initial producer; the other is a cutoff level z_*, so that no bids above z_* are accepted. Note that the cost function (15) is, in general, *not* linear in y inasmuch as F, $\bar{F}$, and $\bar{f}$ depend on y (see equation (11)).

Rather than spell out the general optimality conditions for (15), I consider the following special case. Let $\alpha=1$, G be uniform,

$$G(c)=c, \qquad 0\le c\le 1,$$

$$H(s)=2s, \qquad 0\le s\le 1/2.$$

Then

$$F(z)=1-yz^2, \qquad \bar{F}(z)=yz^2,$$

$$\bar{f}(z)=2yz, \qquad a(z)=z/2.$$

Consequently,

$$(16)\quad TC(z_*,y)$$
$$=ny[1+2y]\int_0^{z_*}(1-yz^2)^{n-1}z^2\,dz$$
$$+z_0(1-yz_*^2)^n.$$

Differentiating (16), we obtain

$$(17a)\quad \partial\{TC\}/\partial z_*$$
$$=ny(1+2y)(1-yz_*^2)^{n-1}z_*^2$$
$$-2nz_0(1-yz_*^2)^{n-1}z_*y,$$

$$(17b)\quad \partial\{TC\}/\partial y$$
$$=n\Big\{(4y+1)\int_0^{z_*}(1-yz^2)^{n-1}z^2\,dz$$
$$-(n-1)y(2y+1)\int_0^{z_*}(1-yz^2)^{n-2}z^4\,dz$$
$$-z_0(1-yz_*^2)^{n-1}z_*^2\Big\}.$$

Integrating the second term in (17b) by parts and collecting terms, one obtains

$$(17a')\quad (1/nyz_*)(\partial\{TC\}/\partial z_*)$$
$$=(1+2y)(1-yz_*^2)^{n-1}z_*$$
$$-2z_0(1-yz_*^2)^{n-1}$$
$$=(1-yz_*^2)^{n-1}[(1+2y)z_*-2z_0],$$

$$(17b')\quad (2/n)(\partial\{TC\}/\partial y)$$
$$=(2y-1)\int_0^{z_*}(1-yz^2)^{n-1}z^2\,dz$$
$$+(1-yz_*^2)^{n-1}z_*^2[(1+2y)z_*-2z_0].$$

Finally, setting these derivatives equal to zero we get $y=1/2$, $z_*=z_0$.

Interpreting these numerical results, note that the winner in the low-bid auction is awarded exactly one-half of the project. In addition, a submission fee is charged so that only contractors with productive efficiency exceeding that of the DOD's fall-back option will enter into the auction. The result of this example runs counter to Optimal Auction Theory where z_* is set strictly below z_0. The

reason for this difference is as follows. In the standard auction (where $y=1$), a transaction is consummated with a single-source contractor. His reservation value, z_i, can be identified then as the cost of the project (including an $R\&D$ component). If, on the other hand, multiple-source contracting is considered, z_i is only the cost to the *initial* contractor; the remainder of the project being competitively purchased at the *lower* cost $a(z_i)$ (assuming $\alpha=1$). The "true" underlying costs are thus below the reservation values z_i and the cutoff point z_* is set higher than it would otherwise be. Clearly, the above figures are special to the example at hand. In general, the design of auction rules will depend on the data of the problem. Explicitly, on what is known about $R\&D$ costs, their distribution across potential producers, the interface efficiency loss (α), and the governmental outside options.

Having stated the governmental cost-minimization problem, I return to the previous investigation. That is, I ask what fraction of the project should be awarded to an initial contractor. In the Appendix, the following closed-form expression for the bid of type z firm is derived:

$$(18) \qquad b(z)=z+\Delta,$$

where the markup, $\Delta>0$, is given by

$$(19) \qquad \Delta(y)=\frac{\int_{z(y)}^{z_*}F^{n-1}(\xi;y)\,d\xi}{F^{n-1}(z;y)}.$$

Using (18) and (19), the governmental objective may be restated as the minimization of the expected value of expression of the form:

$$(20) \qquad \rho(y;z(y,s))=[z(y)+\Delta(y)]\,y + \alpha a(z(y))(1-y).$$

The derivative of (20) with respect to y is computed in Appendix II. It is shown there that

$$(21) \qquad \left.\frac{\partial\rho}{\partial y}\right|_{y=1}=\frac{\int F^{n-1}(g\tilde{G}/G^2)\,d\xi}{F^{n-1}(z)} + \frac{F^{n-1}(z_*)}{F^{n-1}(z)}\frac{\tilde{G}(z_*)}{G(z_*)} + \frac{(n-1)h(\tilde{G})\tilde{G}(z)\int F^{n-1}d\xi}{F^{n}(z)} -(\alpha-1)a(z).$$

Also $\left.\dfrac{d\{TC\}}{dy}\right|_{y=1}=\int\left.\dfrac{\partial\rho}{\partial y}\right|_{y=1}dM(z)$

where $M(z)$ is the cumulative distribution function of $\mathrm{Min}\{z_1,\dots,z_n\}$.

Thus the following *sufficient* condition is obtained.

PROPOSITION 2: *Governmental costs are minimized by initially auctioning only a fraction of the project if*

$$(22) \qquad \frac{\int F^{n-1}(g\tilde{G}/G^2)\,d\xi}{F^{n-1}(z)} + \frac{F^{n-1}(z_*)}{F^{n-1}(z)}\frac{\tilde{G}(z_*)}{G(z_*)} + \frac{(n-1)h\tilde{G}\int F^{n-1}d\xi}{F^{n}(z)} > (\alpha-1)a(z),\ \underline{z}<z<z_*.$$

While condition (22) is not particularly revealing as it stands, two interesting and intuitive consequences are derived from it. First, since its left-hand side is always positive, we see that it always pays to set $y^*<1$ if the interface problem is not too severe; that is, if α is sufficiently close to 1. By way of comparison, recall that under perfect competition we had $y^*=1$ even if switching production from one firm to another entailed no efficiency loss ($\alpha=1$). Second, as I shall

now show, the left-hand side of (22) tends to zero as the number of bidders increases indefinitely. Thus, the sufficient condition (22) is violated for large enough n. This tells us that it would make sense to auction initially a learning buy if the number of competing firms is relatively small. This conclusion is, of course, in conformity with my previous analysis of the perfect competition case. Summarizing, we have

PROPOSITION 3: *Assume that $g\tilde{G}/G^2$ is bounded. A fractional purchase is desirable if either* (a) *the interface efficiency loss is small; or* (b) *the number of potential producers is small.*

PROOF:

Part (a) of the proposition is apparent upon inspection of (22). To prove (b), observe that $F^{n-1}(\xi)/F^{n-1}(z)<1$, $\underline{z}<\xi<z_*$. Thus as $n\to\infty$, both $F^n(\xi)/F^n(z)$ and $nF^n(\xi)/F^n(z)$ tend to zero. By the bounded convergence theorem, the left-hand side of (22) $\to 0$ as $n\to\infty$. Therefore, for sufficiently large n, the inequality is violated and the claim is proven.

Remark: The premise underlying Proposition 3 is not too restrictive. It would be satisfied if, for example, $0<a\leq g(s)\leq A<\infty$.

Since, as already noted, the number of potential producers is typically small, the idea of implementing learning buys seems quite attractive. This conclusion is consistent both with the public sentiment and with longstanding recommendations of the Joint Economic Committee. Empirical studies, summarized by Yuseph, have clearly indicated that 50 percent unit price reductions (at the very least) have been realized by implementing learnings buys. These findings pertain either to moderate size items, common hardware, consumer goods, etc., or to spare parts of major weapon systems. Interestingly, despite the continued pressure, the DOD still spends 60 to 70 percent of its budget purchasing from sole-source producers. Neither bidding nor learning buys have been attempted on a large scale.

III. Variable Size Projects

Up until now I have assumed that the DOD predetermines the number of units it wishes to purchase. The type of benefit function that induces this type of behavior is, of course, special. More realistically, a procuring agent would allow the purchasing decision to depend on the results of the *R&D* stage. This added flexibility would certainly improve performance and it is, in fact, a common practice of the DOD. (Actually, even changes in the design and definition of the product itself are sometimes included.)

My goal in this section is to account for this possibility. This extension is fairly straightforward, so the exposition will be brief. Let $B(x)$ denote the monetary (social) benefit derived from x units of the product. Maintaining the previous notation, the government's *ex post* (having precommitted itself to buy y units from the developer) objective is

$$(23)\qquad \operatorname*{Max}_{x} B(x+y)-\alpha cx.$$

The necessary first-order conditions for this problem are

$$(24\text{a})\qquad B'(x+y)-\alpha c\leq 0,$$

$$(24\text{b})\qquad x[B'(x+y)-\alpha c]=0.$$

Letting $U(c;y)$ be the maximum value function associated with the above program, the expected benefit can be written as

$$(25)\qquad \int_{\underline{z}}^{z_*}\int_0^z U(c;y)\,dG^z(c)\,dF^n(z),$$

where $\quad G^z(c)\equiv G(c)/G(z),0\leq c\leq z.$

From this we must subtract the cost of initially purchasing y units. This cost has already been computed to be

$$(26)\quad ny\int F^{n-1}(z)[\bar{F}(z)+z\bar{f}(z)]\,dz.$$

Finally, the possibility of in-house production must be accounted for. If we let s_0

denote the government's per unit $R\&D$ cost, then expected total benefit (including the $R\&D$ and production phases) is $U(z_0;0)$, where z_0 is defined by

$$s_0 = \int_0^{z_0}[U(c;0) - U(z_0;0)]\,dG(c).$$

(See Appendix III.) Since in-house production occurs with probability $F^n(z_*)$, the expected benefit is

$$(27)\qquad U(z_0;0)F^n(z_*).$$

Adding up these three components the government's objective function becomes

$$(28)\qquad \underset{z_*,y}{\text{Max}} \int_{\underline{z}}^{z_*}\int_0^z U(c;y)\,dG^z(c)\,dF^n(z) + U(z_0;0)F^n(z_*) - ny\int_{\underline{z}}^{z_*}F^{n-1}(z)[\bar{F}(z) + z\bar{f}(z)]\,dz.$$

Since (28) is only a modest modification of (15), its analysis is quite analogous. The details are therefore omitted.

IV. Conclusion

Of the many issues underlying the weapon acquisition process, two seem to be particularly interesting. These are the initial choice of a cost-efficient contractor and the incentives that it has to pursue an effective production plan. In this paper an attempt to simultaneously address these issues has been attempted. I have done so by integrating the theory of optimal auctions with the theory of incentive contracts. The primary conclusion which emerges from the study is that under the conditions which characterize the defense industry (small number of competitors; etc.), it would be expedient to invoke learning buys. This will bring in new suppliers at the later stages of the project, removing thereby the monopolistic shelter under which many contractors are currently protected. While this arrangement has some adverse effects in terms of the initial contractor's incentives to develop a low-cost product and the "interface problem," it still dominates the single-source contractor arrangement.

Appendix

I

Except for the term involving the purchase of the additional $1 - y$ units, the derivations below parallel the arguments in Riley and Samuelson. Starting from (14), I impose the Nash-response constraint that $\pi(z_1; z_1) \geq \pi(x; z_1)$ for all x and z_1. Assuming sufficient differentiability, this leads to the first-order conditions

$$(\text{A1})\qquad P'(z) = z(d/dz)F^{n-1}(z)$$

and the boundary condition

$$(\text{A2})\qquad P(z_*) = z_*F^{n-1}(z_*),$$

where z_* corresponds to the marginal firm entering the auction. Integrating (A1), we get

$$\begin{aligned} P(z_*) - P(z) &= \int_z^{z_*}\xi\,dF^{n-1}(\xi) \\ &= z_*F^{n-1}(z_*) - zF^{n-1}(z) - \int_z^{z_*}F^{n-1}(\xi)\,d\xi \\ &= P(z_*) - zF^{n-1}(z) - \int_z^{z_*}F^{n-1}(\xi)\,d\xi, \end{aligned}$$

where we have substituted from (A2) to derive the last equality.

Therefore,

$$(\text{A3})\qquad P(z) = zF^{n-1}(z) + \int_z^{z_*}F^{n-1}(\xi)\,d\xi.$$

This justifies equation (19). Indeed, by (13):

$$\begin{aligned} b(z) &= P(z)/F^{n-1}(z) \\ &= \frac{zF^{n-1}(z) + \int F^{n-1}(\xi)\,d\xi}{F^{n-1}(z)} \\ &= z + \frac{\int F^{n-1}(\xi)\,d\xi}{F^{n-1}(z)} = z + \Delta. \end{aligned}$$

Returning to the minimization problem, recall that $P(z)$ is the expected payoff to a representative firm of type z. Thus, we must average over all types entering into the bidding game.

$$\int_{\underline{z}}^{z_*} P(z)\bar{f}(z)\,dz$$
$$= \int_{\underline{z}}^{z_*}\left[zF^{n-1}(z)+\int_{z}^{z_*}F^{n-1}(\xi)\,d\xi\right]\bar{f}(z)\,dz$$
$$= \int_{\underline{z}}^{z_*} zF^{n-1}(z)\bar{f}(z)\,dz + \int_{\underline{z}}^{z_*}\int_{z}^{z_*}F^{n-1}(\xi)\,d\xi\,\bar{f}(z)\,dz$$
$$= \int_{\underline{z}}^{z_*} zF^{n-1}(z)\bar{f}(z)\,dz + \int_{\underline{z}}^{z_*}\int_{z}^{z_*}\bar{f}(z)\,dz\,F^{n-1}(\xi)\,d\xi$$
$$= \int_{\underline{z}}^{z_*} zF^{n-1}(z)\bar{f}(z)\,dz + \int_{\underline{z}}^{z_*}\bar{F}(\xi)F^{n-1}(\xi)\,d\xi$$
$$= \int_{\underline{z}}^{z_*} F^{n-1}(z)[\bar{F}(z)+z\bar{f}(z)],$$

where I have, first, substituted from (A3) and then changed the order of integration. Since there are n bidders, the cost of contracting out y units is

$$\text{(A4a)}\quad ny\int_{\underline{z}}^{z_*}F^{n-1}(z)[\bar{F}(z)+z\bar{f}(z)]\,dz.$$

To this must be added the cost of purchasing the remaining $1-y$ units. If we let $M=\operatorname{Min}\{z_i\}_{i=1}^{n}$, then the average cost of those units is $a(M)$ and the density of M is $nF^{n-1}(z)\bar{f}(z)$. Thus the cost of buying additional units upon completion of the initial contract is

$$\text{(A4b)}\quad \alpha n(1-y)\int_{\underline{z}}^{z_*}a(z)F^{n-1}(z)\bar{f}(z)\,dz.$$

Finally, we must reckon the possibility that no firm qualifies to enter the auction (i.e., that none of them is sufficiently efficient), in which case the government must use its fallback option. Expected costs in this case are

$$\text{(A4c)}\qquad z_0F^n(z_*).$$

Summing up (A4a), (A4b), and (A4c), we obtain (15) in the text.

II

Fixing s and differentiating (20) with respect to y, one obtains

$$\text{(A5)}\quad d\rho/dy=\Delta(y)+y\Delta'(y)-(\alpha-1)a(z) - \alpha((1-y)/y)(\tilde{G}(z)/G(z)).$$

Now:

$$\text{(A6)}\quad \Delta'(y)$$
$$=\left[-F^{n-2}\left[F^{n}\frac{\partial z}{\partial y}+(n-1)F\int F^{n-2}h\tilde{G}\,d\xi\right] -(n-1)h\tilde{G}\int F^{n-1}d\xi\right]\Big/\left[F^{2(n-1)}(z;y)\right]\Big|_{z(y)}$$
$$=-\partial z/\partial y+A+B$$

where, for example,

$$\int F^{n-1}h\tilde{G}\,d\xi = \int_{z(y)}^{z_*}F^{n-1}(\xi;y)h(y\tilde{G}(\xi))\tilde{G}(\xi)\,d\xi$$

and so on.

Observe that

$$\text{(A7)}\quad -(n-1)yF^{n-2}(\xi;y)h(y\tilde{G}(\xi))\tilde{G}(\xi) = (\tilde{G}(\xi)/G(\xi))\cdot(\partial/\partial\xi)\{F^{n-1}(\xi;y)\}$$

As a result, we have

$$\Delta(y)-y(\partial z/\partial y)+yA$$
$$=\Delta(y)-y\frac{\partial z}{\partial y}-\frac{(n-1)y\int F^{n-2}h\tilde{G}\,d\xi}{F^{n-1}}$$
$$=\Delta(y)-y(\partial z/\partial y)+$$

$$+\frac{\int_z^{z_*}\frac{\tilde{G}(\xi)}{G(\xi)}\frac{\partial}{\partial\xi}\{F^{n-1}(\xi;y)\}}{F^{n-1}(z)}$$

$$=\Delta(y)-y(\partial z/\partial y)$$

$$+\frac{F^{n-1}(\xi)\frac{\tilde{G}(\xi)}{G(\xi)}\bigg|_z^{z_*}-\int_z^{z_*}F^{n-1}\left(1-\frac{g\tilde{G}}{G^2}\right)d\xi}{F^{n-1}(z)};$$

using (A7), and then integrating by parts. Proceeding,

$$=\Delta(y)-y\frac{\partial z}{\partial y}+\frac{F^{n-1}(z_*)}{F^{n-1}(z)}\frac{\tilde{G}(z_*)}{G(z_*)}$$

$$-\frac{F^{n-1}(z)}{F^{n-1}(z)}\frac{\tilde{G}(z)}{G(z)}-\frac{\int F^{n-1}\left(1-\frac{g\tilde{G}}{G^2}\right)d\xi}{F^{n-1}(z)}$$

$$=\frac{\int F^{n-1}d\xi}{F^{n-1}(z)}+\frac{\tilde{G}(z)}{G(z)}+\frac{F^{n-1}(z_*)}{F^{n-1}(z)}\frac{\tilde{G}(z_*)}{G(z_*)}$$

$$-\frac{\tilde{G}(z)}{G(z)}-\frac{\int F^{n-1}\left(1-\frac{g\tilde{G}}{G^2}\right)d\xi}{F^{n-1}(z)},$$

substituting from (5) and (19). Therefore,

$$\text{(A8)}\quad \partial\rho/\partial y|_{y=1}$$

$$=\Delta(y)+y\Delta'(y)-(\alpha-1)a(z)$$

$$=\Delta(y)-y\frac{\partial z}{\partial y}+yA+yB-(\alpha-1)a(z)$$

$$=\frac{\int F^{n-1}(g\tilde{G}/G^2)d\xi}{F^{n-1}(z)}+\frac{F^{n-1}(z_*)}{F^{n-1}(z)}\frac{\tilde{G}(z_*)}{G(z_*)}$$

$$+\frac{(n-1)h(\tilde{G})\tilde{G}\int F^{n-1}d\xi}{F^n(z)}-(\alpha-1)a(z)$$

which is what we set out to prove.

III

In analogy with equations (1)–(7), let us express the government's functional equation:

$$\text{(A9)}\quad V(n,c_m)$$

$$=\text{Max}\{U(c_m)-ns,\ E_sV(n+1,\ c_m\Lambda c)\}.$$

(The above incorporates both the *R&D* and production phases.)

Define $z(s)$ by

$$s=\int_0^z[U(c)-U(z)]\,dG(c).$$

One can then verify that the following solves equation (A9):

$$\text{(A10)}\quad V(n,c_m)=\begin{cases}U(c_m)-ns & c_m<z(s)\\ U(z)-ns & c_m>z(s)\end{cases}$$

Finally, the *ex ante* total benefit is

$$\text{(A11)}\quad E_cV(1,c)=-s+\int_0^z U(c)\,dG(c)$$

$$+U(z)[1-G(z)]=-s+\int_0^z[U(c)$$

$$-U(z)]\,dG(c)+U(z)=U(z)$$

which justifies equation (27).

REFERENCES

Baron, David P., "Incentive Contracts and Competitive Bidding," *American Economic Review*, June 1972, *62*, 384–94.

Cummins, Michael J., "Incentive Contracting for National Defense: A Problem of Optimal Risk Sharing," *Bell Journal of Economics*, Spring 1977, *8*, 168–85.

Holt, Charles A., "Uncertainty and Bidding for Incentive Contracts," *American Economic Review*, September 1979, *69*, 697–705.

Livne, Zvi and Shubik, Martin, "Naval Procurement Problems: Theory and Practice," Cowles Foundation Discussion Paper No. 627, 1982.

McCall, John J., (1970a) "The Economics of Information and Job Search," *Quarterly Journal of Economics*, February 1970, *84*, 113–26.

______, (1970b) "The Simple Economics of Incentive Contracting," *American Economic Review*, December 1970, *60*, 837–46.

Myerson, Roger, "The Design of Optimal Auctions," *Mathematics of Operations Research*, February 1981, *6*, 58–73.

Nordhaus, William D., *Invention, Growth and Welfare*, Cambridge: MIT Press, 1969.

Peck, M. J. and Scherer, F. M., *The Weapons Acquisition Process — An Economic Analysis*, Cambridge: Harvard University Press, 1962.

Riley, John G. and Samuelson, William, "Optimal Auctions," *American Economic Review*, June 1981, *71*, 321–40.

Scherer, F. M., "The Theory of Contractual Incentives for Cost Reduction," *Quarterly Journal of Economics*, May 1964, *78*, 257–80.

Weitzman, Martin L., "Efficient Incentive Contracts," *Quarterly Journal of Economics*, June 1980, *94*, 719–30.

Williamson, Oliver E., "The Economics of Defense Contracting: Incentives and Performance," in R. N. McKean, ed., *Issues in Defense Economics*, New York: Columbia University Press, 1967, 217–56.

Yuseph, Larry, "A Case for Increasing the Use of Competitive Procurement in the Department of Defense," in Y. Amihud, ed., *Bidding and Auctions for Procurement and Allocation*, New York: New York University Press, 1976.

Executive Office of the President, *Budget of the U.S. Government, FY 1985*, Washington: USGPO, 1984.

A Supergame-Theoretic Model of Price Wars during Booms

By JULIO J. ROTEMBERG AND GARTH SALONER*

This paper explores the response of oligopolies to fluctuations in the demand for their products. In particular, we argue on theoretical grounds that implicitly colluding oligopolies are likely to behave more competitively in periods of high demand. We then show that, in practice, during those periods, various oligopolistic industries tend to have relatively low prices. The few price wars which have been documented also seem to have taken place during periods of high demand. Finally, we study the possibility that this oligopolistic behavior has macroeconomic consequences. We show that it is possible that the increase in competitiveness that results from a shift in demand towards goods produced by oligopolies may be sufficient to raise the output of all sectors.

We examine implicitly colluding oligopolies of the type introduced by James Friedman (1971). These obtain above competitive profits by the threat of reverting to competitive behavior whenever a single firm does not cooperate. This threat is sufficient to induce cooperation by all firms. It must be pointed out that there are usually a multitude of equilibria in such settings. Following Robert Porter (1983a), we concentrate on the best equilibrium of this type the oligopoly can achieve.

The basic point of this paper is that oligopolies find implicit collusion of this type more difficult when their demand is relatively high. The reason for this is simple. When demand is relatively high and price is the strategic variable, the benefit to a single firm from undercutting the price that maximizes joint profits is larger. A firm that lowers its price slightly gets to capture a larger market until the others are able to change their prices. On the other hand, the punishment from deviating is less affected by the state of demand if punishments are meted out in the future, and demand tends to return to its normal level. Thus, when demand is high, the benefit from deviating from the output that maximizes joint profits may exceed the punishment a deviating firm can expect.

What should the oligopoly do when it cannot sustain the level of output that maximizes joint profits? It basically has two alternatives. The first is to give up any attempt to collude when demand is high. This leads to competitive outcomes in booms. Such competitive outcomes are basically price wars. The second, more profitable, alternative is to settle for the highest level of profits (lowest level of output) which is sustainable. As the oligopoly attempts to sustain lower profits, the benefits to a deviating firm fall. Thus, for a given punishment, there is always a level of profits low enough that no single firm finds it profitable to deviate. As demand increases, the oligopoly generally finds that the incentive to deviate is such that it must content itself with outcomes further and further away from those that maximize joint profits.

Our strongest results are for the case in which prices are the strategic variables and marginal costs are constant. Then, increases in demand beyond a certain point actually lower the oligopoly's prices monotonically. This occurs for the following reason: Suppose the oligopoly were to keep its prices constant and only increase output in response to higher demand. Then industry profits would increase when demand goes up. However, in this case, a deviating firm can capture the entire industry profits by shading its price slightly. Therefore, constant prices

*Sloan School of Management and Department of Economics, MIT, Cambridge, MA 02139, respectively. We are grateful to Robert Porter, James Poterba, and Lawrence Summers for many helpful conversations and to two anonymous referees for useful comments. Financial support from the National Science Foundation (grants SES-8209266 and SES-8308782) is gratefully acknowledged.

would increase the incentive to deviate. Reductions in price are needed to maintain implicit collusion.

It might be thought that if firms are capacity constrained in booms, they are essentially unable to deviate, so that the oligopoly doesn't have to cut prices in booms. Indeed, we find that when marginal costs increase with output, a more plausible way of capturing the importance of capacity, our results are weaker. Nonetheless, even in this case the equilibrium can be more competitive when demand is high, whether output or price is the strategic variable.

Any theory whose implication is that competitive behavior is more likely to occur in booms must confront the industrial organization folklore which is that price wars occur in recessions. This view is articulated for example in F. M. Scherer (1980). Our basis for questioning it is not only theoretical. Indeed, it is possible to construct models in which recessions induce price wars.[1,2] In a model with imperfect observability of demand, Edward Green and Porter (1984) show that price wars occur when demand is unexpectedly low. Then, firms switch to competition because they confuse the low price that prevails in equilibrium with cheating on the part of other firms.

Whether competition is more pervasive in booms or busts is an empirical question. While we do not conclusively settle this empirical issue, a brief analysis of some related facts seems to provide more support for our theory than for the industrial organization folklore.

First, at a very general level, it certainly appears that business cycles are related to sluggish adjustment of prices (see Rotemberg, 1982, for example). Prices rise too little in booms and fall too little in recessions. If recessions tended to produce massive price wars, this would be an unlikely finding. Second, more specifically, we find that both Scherer's evidence and our own study of the cyclical properties of price-cost margins are consistent with our theory. The ratio of prices to our measure of marginal cost tends to be countercyclical in more concentrated industries. Also the price wars purported to have happened in the automobile industry (Timothy Bresnahan, 1981) and the railroad industry (Porter, 1983a) occurred in periods of high demand. Finally, since Scherer singles out the cement industry as having repeated breakups of its cartel during recessions, we study the cyclical properties of cement prices. To our surprise, cement prices are strongly countercyclical, even though cement, as construction as a whole, has a procyclical level of output.

Up to this point we have focused on the effect of changes in demand like those that could be induced by business cycles on oligopolistic sectors. We go on to examine whether these oligopolistic responses to changes in demand themselves have aggregate consequences. In particular, we consider the general equilibrium effects of a shift in demand towards an oligopolistic sector. We show that in a very simplified two-sector model, the ensuing reduction in the oligopoly's price can lead the other sector to raise its output as well. This occurs in our model because the other sector, which is competitive, uses the oligopoly's output as an input.

The paper proceeds as follows. Section I presents our theory of oligopoly under fluctuating demand. Section II contains the empirical regularities which lend some plausibility to our theory. Section III considers the general equilibrium model which forms the basis of our discussion of macroeconomics, and conclusions are drawn in Section IV.

[1] If firms find borrowing difficult, recessions might be the ideal occasions for large established firms to elbow out their smaller competitors.

[2] There are also two alternative reasons why prices may be lower when demand is high. First, firms may be charging the monopoly price in the face of short-run increasing returns to scale. The existence of such increasing returns strike us as unlikely. When production is curtailed this is usually done by temporary closings of plants or reductions of hours worked. These reductions would always start with the most inefficient plants and workers thus suggesting at most constant returns to labor in the short run. Second, as argued by Joseph Stiglitz (1984) using a setup similar to the incomplete information limit pricing model of Paul Milgrom and John Roberts (1982), limit pricing may be more salient in booms if the threat of potential entry is also greater at that time.

I. Equilibrium in Oligopolistic Supergames with Demand Fluctuations

We consider N symmetric firms producing a homogeneous good in an infinite-horizon setting. It is well-known that infinitely lived oligopolies of this type are usually able to sustain outcomes in any period that strictly dominate the outcome in the corresponding one-period game, even if firms cannot sign binding contracts. In order to achieve this, the equilibrium strategies must involve a mechanism that deters an individual firm from "cheating" (by expanding output or by shading prices). One such mechanism, and one that has been fruitfully employed in theoretical models,[3] is the use of punishments against the defecting firm in periods following the defection. If these punishments are large enough to outweigh the gain from cheating, then the collusive outcome is sustainable.

In order for the equilibrium strategies to be sequentially rational,[4] however, it must be the case that if a defection actually occurs, the nondefecting firms are willing to mete out the proposed punishment. A simple and often employed way (see Green and Porter, for example) to ensure sequential rationality is for punishments to involve playing the equilibrium strategies from the one-period game for some fixed period of time. We also restrict attention to strategies of this kind. In addition to their simplicity and conformity with the literature, they are also optimal punishments in some cases.[5] The major departure of our model from those that have previously been studied is that we allow for observable shifts in industry demand. We write the inverse demand function as $P(Q_t, \varepsilon_t)$ where Q_t is the industry output in period t and ε_t is the realization at t of $\tilde{\varepsilon}$, the random variable denoting the observable demand shock. We assume that P is increasing in ε_t, that $\tilde{\varepsilon}$ has domain $[\underline{\varepsilon}, \bar{\varepsilon}]$ and a distribution function $F(\varepsilon)$, and that these are the same across periods (i.e., shocks are independently and identically distributed). We denote firm i's output in period t by q_{it} so that

$$Q_t = \sum_{i=1}^{N} q_{it}.$$

The timing of events is as follows: At the beginning of each period, all firms learn the realization of $\tilde{\varepsilon}$ (more precisely ε_t becomes common knowledge). Firms then simultaneously choose the level of their choice variable (price or quantity). These choices then determine the outcome for that period in a way that depends on the choice variable: in the case of quantities, the price clears the market given Q_t; in the case of prices, the firm with the lowest price sells as much as it wants at its quoted price; the firm with the second lowest price then supplies as much of the remaining demand at its quoted price as it wants, and so on. The strategic choices of all the firms then become common knowledge and this one-period game is repeated.

The effect of the observability of ε_t and the key to the difference between the model and its predecessors is the following: the punishments that firms face depend on the future realizations of $\tilde{\varepsilon}$. The expected value of such punishments therefore depends on the expected value of $\tilde{\varepsilon}$. However, the reward for cheating in any period depends on the observable ε_t. We show that for a wide variety of interesting cases, the reward for cheating from the joint profit-maximizing level is monotonically increasing in ε_t. If ε_t is large enough, the temptation to cheat outweighs the punishment.[6] The observability of ε_t allows the oligopoly to recognize this fact. Thus an implicitly colluding oligopoly may

[3] See, for example, Friedman, Green and Porter, and Roy Radner (1980).

[4] Sequentially rational strategies are analyzed in games of incomplete information by David Kreps and Robert Wilson (1982). For the game of complete information that we analyze we use Reinhard Selten's concept of subgame perfection (1965).

[5] When quantities are the strategic variable, Dilip Abreu (1982) shows that punishments can be more severe while still being credible. However, he requires that firms who defect from the punishment be punished in turn, and so on. This considerably complicates the analysis.

[6] In informal discussions, Moses Abramowitz (1938) and Mordecai Kurz (1979) recognize the link between short-run profitability and the sustainability of collusive outcomes. However, the relationship between profits, demand, and costs is not made explicit.

settle on a profit below the fully collusive level in periods of high demand to adequately reduce the temptation to cheat. Such moderation of its behavior tends to lower prices below what they would otherwise be, and may indeed cause them to be lower than for states with lower demand. We illustrate this phenomenon for both the case in which prices and the case in which quantities are the strategic variables.

A. *Prices as Strategic Variables*

We begin with an analysis of the case in which marginal costs (and average costs) are equal to a constant c. This is an appropriate assumption if capacity is very flexible in the short run, if firms produce at under capacity in all states, or if firms produce to order and can accumulate commitments for future deliveries. There always exists an equilibrium in which all the firms set $P = c$ in all periods. Firms then expect future profits to be zero whether they cooperate at time t or not. Accordingly the game at time t is essentially a one-shot game in which the unique equilibrium has all firms setting $P = c$. In what follows we concentrate instead on the equilibria that are optimal for the firms in the industry.

We begin by examining the oligopoly's options for each value of ε_t. Figure 1 shows the profits of each firm, Π, as a function of the aggregate output, Q_t, for a variety of values of ε_t. These profit loci are drawn assuming each firms supplies $1/N$ of Q_t. As ε_t increases, the price for each Q_t rises so that profits are increasing in ε_t. The term $\Pi^m(\varepsilon_t)$ denotes the profit of an individual firm in state ε_t if the firms each produce q^m which equals $1/N$ of the joint profit-maximizing output, Q_t^m. Notice that $\Pi^m(\varepsilon_t)$ is increasing in ε_t since profits are increasing in ε_t even holding Q_t constant.

If a firm deviates from this proposed outcome, it can earn approximately $N\Pi^m$ by cutting its price by an arbitrarily small amount and supplying the entire market demand. Firm i would therefore deviate from the joint profit-maximizing output if

$$(1) \qquad N\Pi^m(\varepsilon_t) - K > \Pi^m(\varepsilon_t)$$

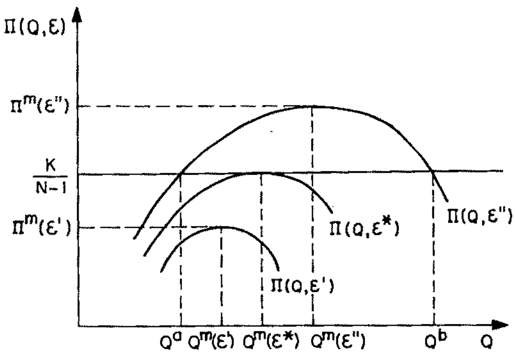


FIGURE 1. PROFITS OF THE OLIGOPOLY

that is, if

$$\Pi^m(\varepsilon_t) > K/(N-1),$$

where K is the punishment inflicted on a firm in the future if it deviates at time t. It is thus the difference between the expected discounted value of profits from $t+1$ on, if the firm goes along, and the expected discounted value of profits if it deviates.

For the moment we will take K to be exogenous and independent of the value of ε_t at the point that cheating occurs. (We will prove the latter shortly and also endogenize K.)

Since $\Pi^m(\varepsilon_t)$ is increasing in ε_t there is some highest level of demand shock, $\varepsilon_t^*(K)$, for which $(N-1)\Pi^m(\varepsilon_t^*) = K$. We consider separately the cases in which ε_t is below and above ε_t^*. In the former cases no individual firm has an incentive to deviate from the joint profit-maximizing outcome. Therefore, if we define $\Pi^s(\varepsilon_t, \varepsilon_t^*)$ to be the highest profits the oligopoly can obtain, $\Pi^s(\varepsilon_t, \varepsilon_t^*) = \Pi^m(\varepsilon_t)$. In the latter case, however, the monopoly profits are not sustainable since any individual firm would have an incentive to cheat. In this case the maximum sustainable profits are given by $(N-1)\Pi^s(\varepsilon_t, \varepsilon_t^*) = K$.

In summary,

$$(2) \quad \Pi^s(\varepsilon_t, \varepsilon_t^*) = \begin{cases} \Pi^m(\varepsilon_t) & \text{for } \varepsilon_t \le \varepsilon_t^* \\ \Pi^m(\varepsilon_t^*) = \dfrac{K}{N-1} & \text{for } \varepsilon_t > \varepsilon_t^*. \end{cases}$$

From (2) it is clear that the sustainable profits are higher, the higher is the punishment. Since we want to concern ourselves with equilibrium strategies that are optimal for the oligopoly, we concentrate on profits that are as large as possible. These involve the lowest possible present discounted value of profits if the firm deviates. Thus charging a price equal to c in all periods following a defection seems optimal, particularly since such punishments never need to be implemented in equilibrium.[7]

However, there are several related reasons why such infinite-length punishments are unlikely to be carried out in practice. First, once the punishment period has begun, the oligopoly would prefer to return to a more collusive arrangement. Second, if the industry members (whether they be firms or even management teams) change over time, shorter punishments seem more compelling. Finally, one can think the reason why firms succeed in punishing each other at all (even though punishments are costly) is because of the anger generated when a rival cheats on the implicit agreement. This anger, as any "irritational" emotion, may be short-lived.

The presence of relatively short punishments is important to our analysis because they make K low. Otherwise the inequality in (1) is always satisfied, that is, in all states of nature the punishment exceeds the benefits from cheating from the collusive price. This is particularly true if the length of the period in which a firm can undercut its competitor's price successfully is short. Thus the inequality in (1) is also more likely to be violated for high ε_t if firms are fairly committed to their current prices as they would be if adjusting prices were costly.

While short periods of punishment are realistic, infinite punishments are simpler. Thus we actually use infinite punishments and capture their relatively small importance by assuming that δ, the factor used to discount future profits, is small.[8] With price equal to marginal cost, the punishment is equal to the discounted present value of profits that the firm would have earned had it not deviated, or

$$(3)\qquad K=\frac{\delta}{1-\delta}\int_{\underline{\varepsilon}}^{\bar{\varepsilon}}\Pi^{s}(\varepsilon,\varepsilon_t^*)\,dF(\varepsilon).$$

Even if we allow K to depend on ε_t, the right-hand side of (2) is independent of ε_t. Therefore the punishment is indeed independent of the state.[9] Using (2) we can rewrite equation (3) as

$$(4)\qquad K(\varepsilon_t^*)=\frac{\delta}{1-\delta}\left[\int_{\underline{\varepsilon}}^{\varepsilon_t^*}\Pi^{m}(\varepsilon)\,dF(\varepsilon)+(1-F(\varepsilon_t^*))\Pi^{m}(\varepsilon_t^*)\right].$$

This gives a mapping from the space of possible punishments into itself: a given punishment implies a cutoff ε_t^* from (2) which in turn implies a new punishment from (4).

The equilibria of the model are the fixed points of this mapping. The equilibrium that is optimal for the oligopoly is the one corresponding to the fixed point with the highest value of K.

It remains to provide sufficient conditions for the existence of a fixed point, that is, to show there exists an $\varepsilon^*\in(\underline{\varepsilon},\bar{\varepsilon})$ for which (2) and (4) hold. Let ε_t' be a candidate for such an ε_t^* and define

$$(5)\qquad g(\varepsilon_t')=\Pi^{m}(\varepsilon_t')-K(\varepsilon_t')/(N-1).$$

We need to show there exists an $\varepsilon_t'\in(\underline{\varepsilon},\bar{\varepsilon})$ such that $g(\varepsilon_t')=0$. Using (4) and (5):

$$g(\underline{\varepsilon})=\Pi^{m}(\underline{\varepsilon})\left(1-\frac{\delta}{(1-\delta)(N-1)}\right)$$

[7]Note that $P=c$ is the highest possible punishment for the oligopoly. If P is below c, firms make losses and will choose not to participate.

[8]An infinite punishment period and low value of δ is only equivalent to a finite punishment period and high value of δ if the length of the punishment is independent of ε_t.

[9]If, instead, the length of the punishment did depend on ε_t, naturally K would depend on ε_t as well.

which is negative if

$$(6) \qquad N < 1/(1-\delta).$$

In other words, for N small enough relative to the discount factor δ, it is possible to obtain the monopoly outcome in at least the lowest state of demand. As N gets bigger, or as firms discount the future more (δ smaller), the punishments become less important and (6) fails.

On the other hand:

$$g(\bar{\varepsilon}) = \Pi^m(\bar{\varepsilon}) - \delta/((N-1)(1-\delta)) \times \int_{\underline{\varepsilon}}^{\bar{\varepsilon}} \Pi^m(\varepsilon)\, dF(\varepsilon)$$

which is positive if

$$(7) \quad \Pi^m(\bar{\varepsilon}) / \int_{\underline{\varepsilon}}^{\bar{\varepsilon}} \Pi^m(\varepsilon)\, dF(\varepsilon) > \delta/[(1-\delta)(N-1)].$$

This condition ensures that the monopoly outcome is not the only solution in every state. This holds when there is sufficient dispersion in the distribution of profit-maximizing outputs. If there is no dispersion, the left-hand side of (7) equals one. Then (7) becomes $N > 1/(1-\delta)$, the opposite of (6). So, in the absence of dispersion, if (6) holds there is never any incentive to cheat. When there is some dispersion, the left-hand side of (7) exceeds one, making it possible for (6) and (7) to hold simultaneously.

If conditions (6) and (7) are satisfied we have: (a) $g(\varepsilon_t')$ is continuous, (b) $g(\bar{\varepsilon}) > 0$, and (c) $g(\underline{\varepsilon}) < 0$, which imply the existence of an $\varepsilon_t' \in (\underline{\varepsilon}_t, \bar{\varepsilon}_t)$ such that $g(\varepsilon_t') = 0$ as required.

This equilibrium has several interesting features. In particular, for $\varepsilon_t > \varepsilon_t^*$ it can be shown that the higher is demand (the higher is ε_t), the higher is equilibrium output and the lower is the equilibrium price. When ε_t exceeds ε_t^*, $\Pi^s = Q_t(P_t - c)$ is constant. Also, Q_t must be as high as possible without reducing firm profits below the sustainable level. In other words, firms must be at Q_t^b in Figure 1 and not at Q_t^a. Otherwise a deviating firm can earn more than $N\Pi^s$ by cutting its price.

Since output is above Q_t^m, profits fall as Q_t rises as can be seen in Figure 1. On the other hand, for a constant Q_t, $Q_t(P_t - c)$ rises as ε_t rises since P_t is larger. Therefore an increase in ε_t must be accompanied by an increase in Q_t. Since increases in ε_t raise profits, increases in Q_t, which lower profits, are required to restore the original level of profits. Moreover, if $Q_t(P_t - c)$ is constant while Q_t rises, P_t must fall. So the oligopoly must actually lower its prices to deter deviations.

The model has some intuitive comparative statics. When N increases and when δ decreases, ε_t^* falls. In both cases, the gains from cheating rise relative to cooperative profits, either because the punishments are distributed among more firms, or because they are discounted more. Thus, the oligopoly must content itself with fewer states in which the monopolistic output is sustained. This can be seen by the following three-part argument.

First, the fact that $g(\bar{\varepsilon})$ is positive ensures that g is increasing in ε at the largest value of ε' for which $g(\varepsilon') = 0$. Second, for fixed Q_t and ε_t, the profits of a single firm are one-Nth of the total profits of the industry. Thus, for a fixed ε_t^*, equation (4) implies that K and $\Pi^m(\varepsilon_t^*)$ are inversely proportional to N. Therefore, increases in N raise g since they raise $\Pi^m(\varepsilon_t^*)$ relative to $K/(N-1)$, that is, the temptation to cheat increases. Similarly, a decrease in δ raises g since K falls. Finally, the increases in g brought about either by an increase in N or a reduction in δ implies that ε_t^* must fall to restore equilibrium.

As mentioned above, punishments are never observed in equilibrium. Thus the oligopoly doesn't fluctuate between periods of cooperation and noncooperation as in the models of Green and Porter. To provide an analogous model, we would have to further restrict the strategy space so that the oligopoly can choose only between the joint monopoly price and the competitive price. Such a restriction is intuitively appealing since the resulting strategies are much simpler and less delicate. With this restriction on strategies, the firms know that when demand is high the monopoly outcome cannot be maintained.

They therefore assume that the competitive outcome will emerge, which is sufficient to fulfill their prophecy. In many states of the world, the oligopoly will earn lower profits than under the optimal scheme we have analyzed. As a result, since punishments are lower, there will be fewer collusive states than before. There will still be some cutoff, ε_t^*, that delineates the cooperative and noncooperative regions. In contrast to the optimal model, however, the graph of price as a function of state will exhibit a sharp decline after ε_t^* with $P = c$ thereafter.

The above models impose no restrictions on the demand function except that it be downward sloping and that demand shocks move it outwards. However, the model does assume constant marginal costs. The case of increasing marginal cost is more complex than that of constant marginal costs for four reasons: 1) A firm that cheats by price cutting does not always want to supply the industry demand at the price it is charging. Specifically, it would never supply an output at which its marginal cost exceeded the price. 2) Cheating now pays off when $\Pi^d(\varepsilon_t, P) > \Pi^s(\varepsilon_t) + K$, where Π^d is the profit to the firm that defects when its opponents charge P. However, Π^d is no longer equal to $(N-1)\Pi^s$. Therefore, the sustainable profit varies by state. 3) With increasing marginal cost, cheating can occur by raising as well as by lowering prices. If its opponents are unwilling to supply all of demand at their quoted price, a defecting firm is able to sell some output at higher prices. 4) The one-shot game with increasing marginal cost does not have an equilibrium in which price is equal to marginal cost. Indeed the only equilibrium is a mixed-strategy equilibrium.[10]

A number of results can nonetheless be demonstrated for an example in which demand and marginal costs are linear:[11]

$$P = a + \varepsilon_t - bQ_t, \tag{8}$$

$$c(q_{it}) = cq_{it} + dq_{it}^2/2. \tag{9}$$

It is straightforward to show that in this example, cheating becomes more desirable as ε_t rises.[12] So, as before, if the oligopoly is restricted to either collude or compete, high ε_t's generate price wars. Alternatively the oligopoly can pick prices P^s which just deter potentially deviating firms. These prices equate Π^s, the profits from going along, with $\Pi^d - K$ where K is the expected present value of Π^s minus the profits obtained when all firms revert to noncooperative behavior.

It is thus possible to calculate the P^s's, the sustainable prices, numerically. For a given value of K one first calculates in which states monopoly is not sustainable. For those states the sustainable price must then be calculated. Since both the sustainable profit, Π^s, and the profit to a deviating firm, Π^d, are quadratic in P^s, this involves solving a quadratic equation. The relevant root is the one that yields the highest value of Π^s that is consistent with the deviating firm planning to meet demand or equating price to marginal cost.

The resulting P^s's then enable us to calculate a new value for K: the one that corresponds to the calculated P^s's.[13] We can thus iterate numerically on K starting with a large number. Since larger values of K induce more cooperation, the first K which is a solution to the iterative procedure is the best equilibrium the oligopoly can enforce with competitive punishments. Figure 2 graphs these equilibrium prices and compares them to the monopoly prices as a function of states for a specific configuration

[10]See Eric Maskin (1984) for a proof that a mixed-strategy equilibrium exists.

[11]In this case an increase in ε_t can directly be interpreted as either a shift outwards in demand or a reduction in c, that part of marginal cost which is independent of q. This results from the fact that the profit functions depend on ε_t only through $(a + \varepsilon_t - c)$.

[12]The proof of this is contained in an appendix, available on request.

[13]In order to do this, however, the profits accruing to firms during the punishment period must be calculated. Rather than attempting to solve for the mixed-strategy equilibria, we used the profits corresponding to price equal to marginal cost. In fact, those profits are lower than in the mixed-strategy equilibrium which means that actual punishments are less severe than we have assumed. However, as we show below, even in that case monopoly is often sustainable only in states of low demand. In any case, the qualitative features of the model are unaffected by this assumption, only the actual value of ε_t^* is affected.

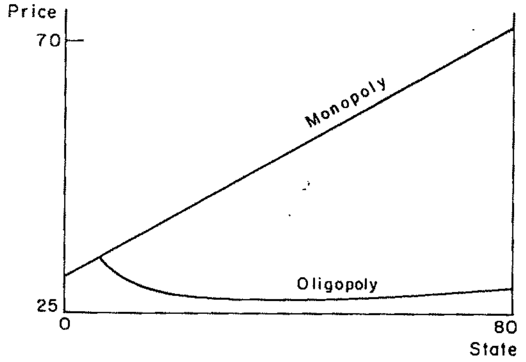


FIGURE 2. PRICES AS STRATEGIC VARIABLES
Parameters: $a=60$, $b=1$, $c=0$, $d=1/3$, $\delta=0.7$, $N=5$

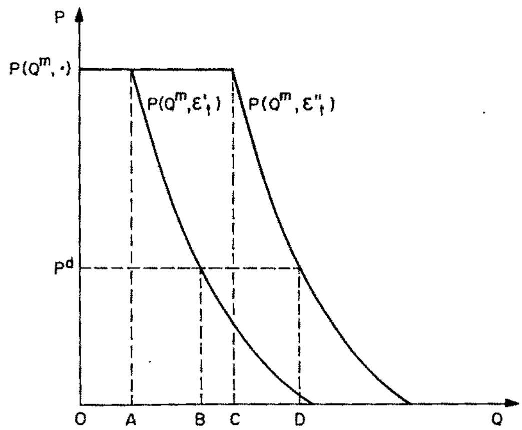


FIGURE 3. THE INCENTIVE TO DEVIATE WITH QUANTITIES AS STRATEGIC VARIABLES

of parameters. In particular ε_t is uniformly distributed over $\{0,1,\dots,80\}$.

As before, the price rises monotonically to ε_t^* and then falls. The major difference here is that eventually the price begins to rise again. The explanation for this is straightforward. In a state with a high value of ε_t, a firm that deviates by shading its price slightly is unwilling to supply all that is demanded at its lower price. Instead, it will supply only to the point where its marginal cost and its price are equal. Now consider such a state and one with slightly more demand. If the oligopoly kept the same price in both states, an individual firm would find that its payoff from deviating is the same in both states (since it would supply to price equals marginal cost in both), but that its profits from going along are higher in the better state. Thus the oligopoly is able to sustain a higher price in the better state.

B. *Quantities as Strategic Variables*

There are two differences between the case in which quantities are used as strategic variables and the case in which prices are. First, when an individual firm considers deviations from the behavior favored by the oligopoly, it assumes that the other firms will keep their quantities constant. The residual demand curve is therefore obtained by shifting the original demand curve to the left by the amount of the rivals' combined output. Second, when firms are punishing each other the outcome in punishment periods is the Cournot equilibrium.

The results we obtain with quantities as strategic variables are somewhat weaker than those we obtained with prices. In particular, it is now not true that any increase in demand (even with constant marginal costs) leads to a bigger incentive to deviate from the collusive level of output. However, we show that when demand and marginal costs are linear, this is the case. We also show with that example that increases in demand can, as before, lead monotonically to "more competitive" behavior.

To see that increases in demand do not necessarily increase the incentive to deviate, we consider the following counterexample. Suppose that demand in states ε_t' and ε_t'' gives rise to the residual demand curves faced by an individual deviating firm in Figure 3. These demand curves are merely horizontal translations by $(N-1)q^m$ of the depicted residual demand curves. The monopoly price, P^m, is the same in both states because there is no demand at prices above P^m. Although these demand curves may seem somewhat contrived, they will suffice to establish a counterexample. They can be rationalized by supposing that there is a substitute good that is perfectly elastically supplied at price P^m.

A deviating firm chooses output to maximize profits given these residual demand curves. Suppose that the maximum profits

are achieved at output D and price P^d for state ε_t''. For this to be a worthwhile deviation, it must be the case that the revenues from the extra sales due to cheating (CD) are greater than the loss in revenues on the old sales from the decrease in price from $P(Q^m, \cdot)$ to P^d. But (except for a horizontal translation) the firm faces the same residual demand curve in both states. Thus by selling at P^d, the extra sales due to cheating are the same at $\varepsilon_t'(AB)$ as at $\varepsilon_t''(CD)$. Moreover the loss in revenue on old sales is strictly smaller at ε_t'. Therefore the firm has a strictly greater incentive to deviate in state ε_t' than in state ε_t''.

The above counterexample exploits the assumed structure of demand only to establish that the collusive price is the same in both states. We have therefore also proved a related proposition: for any demand function, if the oligopoly keeps its price constant when ε_t increases (thus supplying all the increased demand), the incentive to cheat is reduced when demand shifts horizontally. This is why the oligopoly is always able to increase the price as the state improves.

Now consider the case in which demand and marginal costs are linear as in (8) and (9). There an increase in ε_t always leads to a bigger incentive to deviate from the collusive output.[14] As in the previous subsection, if the only options for the oligopoly are to either compete or collude, price wars emerge when demand is sufficiently high. Alternatively, the oligopoly can choose a level of output that will just deter firms from deviating when demand is high. The equilibrium levels of output can be obtained numerically in a manner analagous to the one used to calculate the equilibrium sustainable prices in the previous subsection.

Figure 4 plots the ratio of this equilibrium price to the monopoly price as a function of ε_t. While the equilibrium price rises as ε_t rises, it can be seen that beyond a certain ε_t the ratio of equilibrium price to monopoly price falls monotonically.

[14]The proof of this is also contained in the appendix available on request.

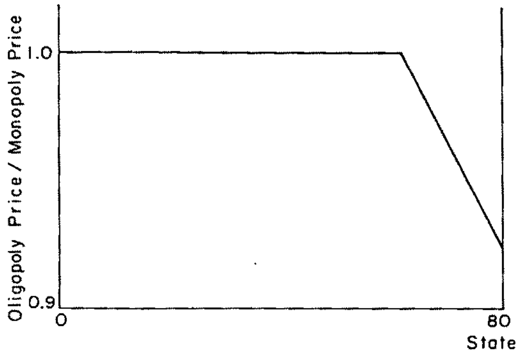


FIGURE 4. QUANTITIES AS STRATEGIC VARIABLES
Parameters: $a = 60$, $b = 1$, $c = 0$, $d = 1/3$, $\delta = 0.7$, $N = 5$

II. A Survey of Related Empirical Findings

The theory presented in the previous section runs counter to the industrial organization folklore. This folklore is best articulated in Scherer, who says: "Yet it is precisely when business conditions really turn sour that price cutting runs most rampant among oligopolists with high fixed costs" (p. 208).

Given the pervasiveness of this folklore, it is incumbent upon us to at least provide some fragments of evidence which are consistent with our theory. There are at least three kinds of data capable of shedding light on whether prices tend to be low in concentrated industries when their demand is high. First, there is the cyclical pattern of prices in concentrated industries relative to other prices. We can see whether these relative prices tend to be pro- or countercyclical. Second, a similar analysis can be applied to the cyclical pattern of prices in concentrated industries relative to their costs. Finally, there are the documented episodes of price wars. Here what is relevant is whether they occurred in periods of high or low demand. In this section, we reexamine existing data of all three types. It must be pointed out at the outset, however, that this analysis is not a direct empirical test of the model itself, but only a cursory analysis of its most striking implication. The need for such direct tests is suggested by our findings since they largely bear out this implication.

TABLE 1—THE CYCLICAL PROPERTIES OF CEMENT PRICES
(Yearly Data from 1947 to 1981)[a]

	Dependent Variable			
Coefficient	P^c/PPI	P^c/PPI	P^c/P^{con}	P^c/P^{con}
Constant	.025 (.010)	.025 (.012)	.038 (.007)	.037 (.008)
GNP	−.438 (.236)	−.456 (.197)	−.875 (.161)	−.876 (.149)
ρ		.464 (.173)		.315 (.183)
R^2	.10	.15	.48	.52
D-W	1.03	1.73	1.28	1.92

[a] P^c is the price of cement, PPI is the Producer Price Index, and P^{con} is the price index of construction materials. Standard errors are shown in parentheses.

A. *The Cyclical Properties of Cement Prices*

Scherer cites three industries whose experience is presented as supporting the folklore: rayon, cement and steel. For rayon he cites a study by Jesse Markham (1952) which shows mainly that the nominal price of rayon fell during the Great Depression. Since broad price indices fell during this period this is hardly proof of a price war. Rayon has since been replaced by other materials making it difficult to use postwar data to check whether any real price-cutting took place during postwar recessions. For steel Scherer admits the following: "...up to 1968 and except for some episodes during the 1929–38 depression, it was more successful than either cement or rayon in avoiding widespread price deterioration, even when operating at less than 65% of capacity between 1958 and 1962" (p. 210).

This leaves cement. We study the cyclical properties of real cement prices below. We collected data on the average price of portland cement from the *Minerals Yearbook* (Bureau of Mines). We then compare this price with the Producer Price Index and the price index of construction materials published by the Bureau of Labor Statistics. Regressions of the yearly rate of growth of real cement prices on the contemporaneous rate of growth of *GNP* are reported in Table 1.

As the table shows, the coefficient of the rate of growth of *GNP* is always meaningfully negative. A 1 percent increase in the rate of growth of *GNP* leads to a 0.5–1.0 percent fall in the price of cement. To test whether the coefficients are significant, the regression equations must be quasi differenced since their Durbin-Watson (*D-W*) statistics are small. Once this is done we find the coefficients are all significantly different from zero at the 5 percent level. More casually, the price of cement relative to the index of construction prices rose in the recession year 1954, while it fell in the boom year 1955. Similarly, it rose during the recession year 1958 and fell in 1959. These results show uniformly that the price of cement has a tendency to move countercyclically as our theory predicts for an oligopoly.

These results are of course not conclusive. First, it is possible that increases in *GNP* lower the demand for cement relative to that for other goods. Without a structural model, which is well beyond the scope of this paper, this question cannot be completely settled. However, the rate of growth of the output of the cement industry has a correlation of .69 with the rate of growth of *GNP*, and of .77 with the rate of growth of construction activity which is well known to be procyclical. Second, our regressions do not include all the variables one would expect to see in a reduced form. Thus the effect of *GNP* might be proxying for an excluded variable like the capacity of cement mines. This variable would probably be expected to exercise a negative effect on the real price of cement. It

must be pointed out, however, that capacity itself is an endogenous variable which also responds to demand. It would thus be surprising if enough capacity were built in a boom to more than offset the increase in demand. If anything, the presence of costs of adjusting capacity would make capacity relatively unresponsive to increases in *GNP*.

B. *The Cyclical Properties of Price-Cost Margins*

In the industrial organization literature there have been a number of studies that have attempted to measure the cyclical variations in price-cost margins. Usually these are measured by sales minus payroll and material costs divided by sales. This is a crude approximation to the Lerner Index which has the advantage of being easy to compute. Indeed, Scherer cites a number of studies which analyzed the cyclical variability of these margins in different industries. These studies have led to somewhat mixed conclusions. However, Scherer concludes: "The weight of the available statistical evidence suggests that concentrated industries do exhibit somewhat different pricing propensities over time than their atomistic counterparts. They reduce prices (and more importantly) price-cost margins by less in response to a demand slump and increase them by less in the boom phase" (p. 357). This does not fit well with the folklore which would predict that, on average, prices would tend to fall more in recessions the more concentrated is the industry. On the other hand, a recent paper by Ian Domowitz, Glenn Hubbard, and Bruce Petersen (1986a) finds more procyclical movements of price-cost margins in concentrated industries.

Price-cost margins can only be interpreted as the Lerner Index if labor costs are proportional to output. However, there is a large fixed component to labor costs. Thus when output rises, the ratio of labor costs to revenues falls and, *ceteris paribus*, price-cost margins rise. Therefore, if the fixed labor cost tends to be higher in concentrated industries, one expects to find their price-cost margins to be relatively procyclical.

We therefore also study some independent evidence on margins. Michael Burda (1984) reports correlations between employment and real product wages in various 2-digit industries. These real product wages are given by the average hourly wage paid by the industry divided by the value-added deflator for the industry. They can be interpreted as a different crude measure of marginal cost over prices. Their disadvantage over the traditional price-cost margin is that, unlike the latter, to interpret them in this way requires not only that materials be proportional to output, but also that materials costs be simply passed through as they would in a competitive industry with this cost structure. On the other hand, their advantage over the traditional measure is that they remain valid when some of the payroll expenditure is a fixed cost as long as, at the margin, labor has a constant marginal product. Moreover, it turns out that if the marginal product of labor actually falls as employment rises, our evidence provides even stronger support for our theory.

The correlations reported by Burda for the real product wage and employment using detrended yearly data from 1947 to 1978 are reported in Table 2, which also reports the average four-firm concentration ratio for each 2-digit industry. This average is obtained by weighting each 4-digit SIC code industry within a particular 2-digit SIC code industry by its sales in 1967. These weights were then applied to the 1967 four-firm concentration indices for each 4-digit SIC code industry obtained from the Census.[15]

At first glance it is clear from Table 2 that more concentrated industries like motor vehicles and electrical machinery tend to have positive correlations while less concentrated industries like leather, food, and wood products tend to have negative correlations. Statistical testing of this correlation with the

[15] When constructing these aggregate concentration indices we systematically neglected the 4-digit SIC code industries which ended in 99. These contain miscellaneous or "not classified elsewhere" items whose concentration index does not measure market power in a relatively homogeneous market.

TABLE 2—CONCENTRATION AND THE CORRELATION BETWEEN REAL WAGES AND EMPLOYMENT

SIC Number	Industry Designation	Corre-lation	Concen-tration
Durables Manufacturing			
24	Lumber and Wood Products	−.33	17.6
25	Furniture and Fixtures	−.18	21.6
32	Stone, Clay and Glass	.39	37.4
33	Primary Metals	.32	42.9
34	Fabricated Metal Industries	.23	29.1
35	Machinery except Electrical	.12	36.3
36	Electrical and Electronic Equipment	.34	45.0
371	Motor Vehicles and Equipment	.19	80.8
372–9	Other Transportation Equipment	.02	50.1
38	Instruments and Related Products	−.36	47.8
Nondurables Manufacturing			
20	Food and Kindred Products	−.30	34.5
21	Tobacco Manufactures	−.64	73.6
22	Textile Mill Products	.04	34.1
23	Apparel and Related Products	−.53	19.7
26	Paper and Allied Products	−.42	31.2
27	Printing and Publishing	.40	18.9
28	Chemical and Allied Products	−.03	49.9
29	Petroleum and Coal Products	−.48	32.9
30	Rubber	.16	69.1
31	Leather and Leather Products	−.44	24.5

concentration index is, however, somewhat delicate. That is because our theory does not predict that an industry which is 5 percent more concentrated than another will reduce prices more severely in a boom. On the contrary, a fully fledged monopoly will always charge the monopoly price which usually increases when demand increases. All our theory says is, that as soon as an industry becomes an oligopoly it becomes likely that it will cut prices in booms.

Naturally the concentration index is not a perfect measure of whether an industry is an oligopoly. Indeed, printing has a low concentration index even though its large components are newspapers, books, and magazines that are in fact highly concentrated, once location in space or type is taken into account. Nonetheless, higher concentration indices are at least indicators of a smaller number of important sellers. Glass is undoubtedly a more oligopolistic industry than shoes. So we classify the sample into relatively unconcentrated and relatively concentrated and choose, somewhat arbitrarily, as the dividing line the median concentration of 35.4. This lies between food and nonelectrical machinery. Table 3 is the resulting 2×2 contingency table.

TABLE 3—CONCENTRATION/CORRELATION CONTINGENCY TABLE

	Unconcentrated	Concentrated	Total
Negatively Correlated	7	3	10
Positively Correlated	3	7	10
Total	10	10	20

An alternative table can be obtained by neglecting the three observations whose correlations are effectively zero. These are sectors 22, 28, and 372–9. Their correlations are at most equal in absolute value to one-third of the next lowest correlation. Then the contingency table has, instead of the values 7:3:3:7, the values 7:2:2:6.

It is now natural to test whether concentrated and unconcentrated industries have the same ratio of positive correlations to negative ones against the alternative that this ratio is significantly higher for concentrated industries. The χ^2 test of independence actually only tests whether the values are unusual under the hypothesis of independence without focusing on our particular alternative. It rejects the hypothesis of independence with 92 percent confidence using the values of Table 3 and with 97 percent confidence using the values 7:2:2:6. This test is, however, likely to be flawed for the small sample we consider. Fisher's test would appear more appropriate since it is an exact test against the alternative that more concentrated sectors have more positive correlations. With

this test the hypothesis that the ratio of positive correlations is the same can be rejected with 91 percent confidence using the data of Table 3 and with 96 percent confidence using 7:2:2:6.[16]

These regularities should be contrasted to the predictions of the standard theory of labor demand. In this theory, employment rises only when the real product wage falls. This occurs in both monopolistic and competitive industries as long as there are diminishing returns to labor. Therefore, the finding that the product wage rises when employment rises suggests the widespread price cutting our theory implies.

There is an alternative classical explanation for our findings. This explanation relies on technological shocks. These shocks can, in principle, either increase or decrease the demand for labor by a particular sector. If they increase the demand and the sector faces an upward-sloping labor supply function, employment and real wages can both increase. The difficulty with this alternative explanation is that the sectors with positive correlations do not appear to be those which a casual observer would characterize as having many technological shocks of this type. In particular, stone, clay and glass, printing and publishing, and rubber appear to be sectors with fairly stagnant technologies. On the other hand, instruments and chemicals may well be among those whose technology has been changing the fastest.

C. *Actual Price Wars*

There have been two recent studies showing that some industries alternate between cooperative and noncooperative behavior. The first is due to Bresnahan (1981). He studies the automobile industry in 1954, 1955, and 1956, and attempts to evaluate the different interpretations of the events of 1955. That year production of automobiles climbed by 45 percent only to fall 44 percent the following year. Bresnahan formally models the automobile industry as choosing prices each year for a given set of models offered by each firm. He concludes that the competitive model of pricing fits the 1955 data taken by themselves while the collusive model fits the 1954 and 1956 data. Those two years exhibited at best sluggish *GNP* growth. *GNP* fell 1 percent in 1954 while it rose 2 percent in 1956. Instead, 1955 was a genuine boom with *GNP* growing 7 percent.[17] Insofar as cartels can only sustain either competitive or collusive outcomes, this is what our theory predicts. Indeed, in our model, the competitive outcomes will be observed only in booms.

Porter (1983b) studies the railroad cartel which operated in the 1880's on the Chicago-New York route. He uses time-series evidence to show that some weeks were collusive while others were not.

We present some of his findings in the first three columns of Table 4. The first column shows an index of cartel nonadherence estimated by Porter. He shows that this index parallels quite closely the discussions in the *Railway Review* and in the *Chicago Tribune* which are reported by Thomas Ulen (1978). The second column reports rail shipments of wheat from Chicago to New York. The third column shows the percentage of wheat shipped by rail from Chicago relative to the wheat shipped by both lake and rail. The fourth column presents the national production of grains estimated by the Department of Agriculture. Finally the last column represents the number of days between April 1 and December 31 that the Straits of Mackinac remained closed to navigation. (They were always closed between January 1 and March 31.)

The three years in which the most severe price wars occurred were 1881, 1884, and 1885. Those are also the years in which rail shipments are the largest, both in absolute terms and relative to lake shipments. This

[16]These results are consistent with evidence by Domowitz, Hubbard, and Petersen (1986b) which shows that value-added deflators tend to be more countercyclical in concentrated industries.

[17]It must be noted that the focus of Bresnahan's study is the 1955 model year which doesn't coincide with the calendar year. Nonetheless his data on prices correspond to April 1955. By that time the boom was well under way.

TABLE 4—RAILROADS IN THE 1880'S

	Estimated Nonadherence	Rail Shipments (Million bushels)	Fraction Shipped by Rail	Total Grain Production (Billion Tons)[a,b]	Days Lakes Closed 4/1–12/31[a]
1880	0.00	4.73	22.1	2.70	35
1881	0.44	7.68	50.0	2.05	69
1882	0.21	2.39	13.8	2.69	35
1883	0.00	2.59	26.8	2.62	58
1884	0.40	5.90	34.0	2.98	58
1885	0.67	5.12	48.5	3.00	61
1886	0.06	2.21	17.4	2.83	50

[a]Obtained from the Chicago Board of Trade (1880–86).
[b]This total is constructed by adding the productions of wheat, corn, rye, oats, and barley in tons.

certainly does not suggest that these wars occurred in periods of depressed demand. However, shipments may have been high only because the railroads were competing even though demand was low. To analyze this possibility, we report the values of two natural determinants of demand. The first is the length of time during which the lakes were closed. The longer the lakes remained closed, the larger was the demand for rail transport. The lakes were closed the longest in 1881 and 1885. These are also the years in which the index of cartel nonadherence is highest. In 1883 and 1884, the lakes remained closed only slightly less time than in 1885 and yet there were price wars only in 1884. The second natural determinant of demand, total grain production, readily explains the anomalous behavior of 1883. In 1883, total grain production was the second lowest in the entire period and in particular, was 12 percent lower than in 1884. This might have depressed demand so much that, in spite of the lake closings, total demand for rail transport was low enough to warrant cooperation.[18]

In summary, the years in which the cartel was unable to collude effectively were also years in which demand seems to have been high.

III. General Equilibrium Consequences

So far we have considered only the behavior of an oligopoly in isolation. To study the aggregate consequences of this behavior, we need to model the rest of the economy. We consider a two-sector general equilibrium model in which the first sector is competitive while the second is oligopolistic. There is also a competitive labor market. To keep the model simple, it is assumed that workers have a horizontal supply of labor at a wage equal to P_{1t}, the price of the competitive good. Since the model is homogeneous of degree zero in prices, the wage itself can be normalized to equal one. So the price of the good produced competitively must also equal one. This good can be produced with various combinations of labor and good 2. In particular the industrywide production function of good 1 is given by

$$(10)\quad Q_{1t} = \alpha Q_{21t} - \frac{\beta Q_{21t}^2}{2} + \gamma L_{1t} - \frac{\xi L_{1t}^2}{2}$$

[18]Our analysis uses annual aggregates rather than the weekly data used by Porter. As the estimate of cartel nonadherence in Table 2 shows, however, the price wars in 1881, 1884, and 1885 did not last the entire year. Indeed, in each of those years there were at least two separate episodes of price wars. Using only annual data we are unable to show that each of the price wars occurred during a high demand period. Some relevant evidence is provided in a more recent study by Porter (1985). There, using weekly data, he finds that price wars were more likely to occur in any period the larger the quantity sold in the previous period. This suggests that price wars tended to begin when firms expected unusually high demand.

where Q_{1t} is the output of the competitive sector at t, Q_{21t} is the amount of good 2 employed in the production of good 1 at t and L_{1t} is the amount of labor used in the production of good 1. Since the sector is competitive the price of each factor and its marginal revenue product are equated. Thus:

$$L_{1t} = (\gamma - 1)/\xi, \tag{11}$$

$$P_{2t} = \alpha - \beta Q_{21t}. \tag{12}$$

On the other hand the demand for good 2 by consumers is given by

$$P_{2t} = n - mQ_{2ct} + e_t,$$

where Q_{2ct} is the quantity of good 2 purchased by consumers, n and m are parameters, and e_t is an independently and identically distributed random variable. Therefore total demand for good 2 is given by

$$\begin{aligned} P_{2t} &= a + \varepsilon_t - bQ_{2t}, \\ a &= (n\beta + m\alpha)/(m+\beta), \\ \varepsilon_t &= e_t\beta/(m+\beta), \\ b &= m\beta/(m+\beta). \end{aligned} \tag{13}$$

Note that equation (13) is identical to equation (8). To continue the parallel with our sections on partial equilibrium, we assume that the labor requirement to produce Q_{2t} is

$$L_{2t} = cQ_{2t} + (d/2)Q_{2t}^2,$$

which implies that, as before, marginal cost is $c + dQ_{2t}$. The model would be unaffected if good 1 were also an input into good 2 since P_{1t} is always equal to the wage. If sector 2 behaved competitively marginal cost would equal P_{2t}. Then output of good 2 would be Q_{2t}^c while price would be P_{2t}^c:

$$Q_{2t}^c = (a + \varepsilon_t - c)/(b+d),$$

$$P_{2t}^c = ((a+\varepsilon_t)d + bc)/(b+d).$$

An increase in ε_t raises both the competitive price and the competitive quantity of good 2. By (12), less of good 2 will be used in the production of good 1 thus leading to a fall in the output of good 1. So, a shift in tastes raises the output of one good and lowers that of the other. The economy implicitly has, given people's desire for leisure, a production possibility frontier.

Similarly, if sector 2 always behaves like a monopolist, increases in ε_t raise both P_{2t} and Q_{2t} thus lowering Q_{1t}. Once again shifts in demand are unable to change the levels of both outputs in the same direction. On the other hand, if the industry behaves like the oligopoly considered in the previous sections, an increase in ε_t can easily lead to a fall in the relative price of good 2.[19] This occurs in three out of the four scenarios considered in Section I. It occurs when the unsustainability of monopoly leads to competitive outcomes whether the strategic variable is price or output as long as increases in ε_t make monopoly harder to sustain. It also always occurs when the strategic variable is prices and the oligopoly plays an optimal supergame. The decrease in P_{2t} in turn leads firms in the first sector to demand more of good 2 as an input and to increase their output. So, a shift in demand towards the oligopolistic goods raises all outputs much as all outputs move together during business cycles.[20]

A number of comments deserve to be made about this model. First, our assumption that the real wage in terms of good 1 is constant does not play an important role. In equilibrium the reduction in P_{2t} raises real wages thus inducing workers to work more even if they have an upward-sloping supply sched-

[19] This fall in the price of a good in response to an increase in its demand would also characterize industries with increasing returns to scale which, for some reason, equated price to average costs.

[20] Business cycles are persistent and thus cannot adequately be modeled as resulting from the independently and identically distributed shifts considered in previous sections. However, what is necessary for prices to be low when demand is high is only that the punishments for deviating be carried out mostly in states of lower demand. This is likely to happen even if demand follows a fairly general stationary process.

ule for labor. Whether this increased supply of labor would be sufficient to meet the increased demand for employees by sector 2 in unclear. If it wasn't, the wage would have to rise in terms of good 1. More interestingly, if the increased supply of labor was large, P_{1t} would have to rise thus increasing employment also in sector 1. This would lead to an expansion even if good 2 was not an input into good 1. This pattern of price movements is consistent with the evidence on the correlation between product wages and employment presented in Section II.

Second, the model can easily be made consistent with the procyclical variation of profits. Even though sector 2 reduces the margin between price and marginal cost as output expands, the difference between revenues and total costs can increase as long as there are fixed costs.

Third, the analysis leaves unexplained the causes of the shifts in sectoral demands. To make sense of actual business cycles, within the context of the models described here, one would have to relate these shifts in demand to changes in the money supply and interest rates which are highly correlated with cyclical fluctuations. While the connection between financial variables and shifts in demand is beyond the scope of this paper, it must be noted that such shifts form part of the popular discussion of the early stages of recoveries. At that point, consumers' desires for cars and other durables picks up.

Our model exhibits a variety of somewhat Keynesian features. First, changes in aggregate output are related to fluctuations in demand and not, unlike in classical models, to changes in supply conditions such as productivity or labor supply.[21] Second, the model has the potential for providing an explanation for the stickiness of prices discussed, for example, in Rotemberg (1982). Suppose that increases in ε_t are correlated with increases in the money supply. Then increases in output are correlated with increases in the money supply. As long as increases in output raise the demand for real money balances, increases in the money supply will be correlated with increases in real money balances. Prices do not rise equiproportionately. Third, we can discuss the multiplier in the context of our model. This concept reflects the idea that increases in demand lead output to rise which then leads to further increases in demand. Here a shift in demand towards an oligopolistic sector can raise that sector's output, lower its prices and thus raise national income. In turn, this increased national income can lead to increases in the demand for other goods produced in other oligopolistic markets, thus lowering their prices and raising their output as well.

IV. Conclusions

The data we study show moderate support for the theories developed in this paper. This suggests that both the theories and their empirical validation deserve to be extended.

The theory of oligopoly might be extended to include also imperfectly observable demand shifts, prices and outputs of the type studies by Green and Porter. The advantage of introducing unobservable shifts in demand is that these can induce reversions to punishing behavior even when all firms are acting collusively. A natural question to ask is whether reversions to punishing behavior that result from unobservable shocks are more likely when everybody expects the demand curve to have shifted out. Unfortunately, this appears to be a very difficult question to answer. Even the features of the optimal supergame without observable shocks discussed in Porter (1983a) are hard to characterize. Adding the complication that both the length of the punishment period as well as the price that triggers a reversion depend on observable demand is a formidable task.

[21] Keynesian models usually focus on changes in "aggregate demand" whereas our model hinges on changes in relative demand. However, in practice, when households demand more, they demand disproportionately more from certain oligopolistic sectors such as the consumer durables sector. Therefore, the distinction between the two types of changes in demand may not be very important.

In this paper we considered only business cycles that are due to the tendency of oligopolists to act more competitively when demand shifts towards their products. An alternative and commonly held view is that business cycles are due to changes in aggregate demand which do not get reflected in nominal wages. In that case, a decrease in aggregate demand raises real wages, thereby reducing all outputs. In our theory of oligopoly, firms tend to collude more in these periods. Hence recessions are not only bad because output is low, but also because microeconomic distortions are greater. This suggests that stabilization of output at a high level is desirable because it reduces these distortions.

On the other hand, the business cycles discussed here do not necessarily warrant stabilization policy. While models of real business cycles merely feature ineffective stabilization policies, here such policies might actually be harmful. Booms occur because, occasionally, demand shifts towards oligopolistic products. In these periods the incentive to deviate from the collusive outcome is greatest, because the punishment will be felt in periods that, on average, have lower demand and hence lower profits. If, instead, future demand were also known to be high, the threat of losing the monopoly profits in those good periods might well be enough to induce the members of the oligopoly to collude now. So, if demand for the goods produced by oligopolies were stable they might collude always, leaving the economy in a permanent recession.[22] Therefore the merits of stabilization policy hinge crucially on whether business cycles are due to shifts in demand unaccompanied by nominal rigidities, or whether they are due to changes in aggregate demand accompanied by such rigidities. Disentangling the nature of the shifts in the demand faced by oligolopies therefore seems to be a promising line of research.

[22] For the examples in Figures 3 and 4, this occurs as long as $\delta \geq 0.8$ when prices are the strategic variables, or $\delta \geq 0.25$ when quantities are the strategic variables.

Much work also remains to be done empirically validating our model itself. In Section II we presented a variety of simple tests capable of discriminating between the industrial organization folklore and our theory. Since none of them favored the folklore, it may well be without empirical content. On the other hand, our theory deserves to be tested more severely. First, a more disaggregated study of the cyclical properties of price-cost margins seems warranted. Unfortunately, data on value-added deflators do not appear to exist at a more disaggregated level so a different methodology will have to be employed. Second, our theory has strong implications for the behavior of structural models of specific industries. The study of such models ought to shed light on the extent to which observable shifts in demand affect the degree of collusion.

Finally, our theory can usefully be applied to other settings. Consider, in particular, the game between countries as they set their tariffs. In standard models, unilateral tariffs may be desirable either as devices to exercise monopsony power or, with fixed exchange rates, to increase employment. The noncooperative outcome in a game between the countries may have very little international trade. In a repeated game, more international trade can be sustained by the threat to curtail trade further. If unilateral trade barriers become more attractive in recessions (because the gains in employment they induce are valued more), the equilibrium will have trade wars in states of depressed demand.

REFERENCES

Abramowitz, Moses, "Monopolistic Selling in a Changing Economy," *Quarterly Journal of Economics*, February 1938, *52*, 191–214.

Abreu, Dilip, "Repeated Games with Discounting: A General Theory and an Application to Oligopoly," mimeo., Princeton University, December 1982.

Bresnahan, Timothy F., "Competition and Collusion in the American Automobile Industry: The 1955 Price War," mimeo., Stanford University, February 1981.

Burda, Michael C., "Dynamic Labor Demand Schedules Reconsidered: A Sectoral Approach," mimeo., Harvard University, June 1984.

Domowitz, Ian, Hubbard, R. Glenn and Petersen, Bruce, (1986a) "Business Cycles and the Relationship between Concentration and Price-Cost Margins," *Rand Journal of Economics*, Spring 1986, *17*, 1–17.

______, ______, and ______, (1986b) "Cyclical Variation in Industry Price-Cost Margins: Price and Cost Channels," mimeo., Northwestern University, 1986.

Friedman, James W., "A Non-Cooperative Equilibrium for Supergames," *Review of Economic Studies*, January 1971, *38*, 1–12.

Green, Edward J. and Porter, Robert H., "Noncooperative Collusion under Imperfect Price Information," *Econometrica*, January 1984, *52*, 87–100.

Kreps, David and Wilson, Robert, "Sequential Equilibrium," *Econometrica*, July 1982, *50*, 863–94.

Kurz, Mordecai, "A Strategic Theory of Inflation," IMSSS Technical Report No. 283, April 1979.

Markham, Jesse W., *Competition in the Rayon Industry*, Cambridge: Harvard University Press, 1952.

Maskin, Eric, "Equilibrium with Price-Setting Duopolists," mimeo., MIT, 1984.

Milgrom, Paul and Roberts, John, "Limit Pricing and Entry under Incomplete Information: An Equilibrium Analysis," *Econometrica*, March 1982, *50*, 443–60.

Porter, Robert H., (1983a) "Optimal Cartel Trigger-Price Strategies," *Journal of Economic Theory*, April 1983, *29*, 313–38.

______, (1983b) "A Study of Cartel Stability: The Joint Economic Committee, 1880–1886," *Bell Journal of Economics*, Autumn 1983, *14*, 301–14.

______,, "On the Incidence and Duration of Price Wars," *Journal of Industrial Economics*, June 1985, *33*, 415–26.

Radner, Roy, "Collusive Behavior in Noncooperative Epsilon-Equilibria of Oligopolies with Long but Finite Lives," *Journal of Economic Theory*, April 1980, *22*, 136–54.

Rotemberg, Julio J., "Sticky Prices in the United States," *Journal of Political Economy*, December 1982, *90*, 1187–211.

Scherer, F. M., *Industrial Market Structure and Economic Performance*, 2nd ed., Boston: Houghton-Mifflin, 1980.

Selten, Reinhard, "Spieltheoretische Behandlung eines Oligopolsmodells mit Nachfrageträgheit," *Zeitschrift für die Gesamte Staatswissenschaft*, 1965, *121*, 301–24 and 667–89.

Stiglitz, Joseph E., "Price Rigidities and Market Structure," *American Economic Review Proceedings*, May 1984, *74*, 350–55.

Ulen, Thomas S., "Cartels and Regulation," unpublished doctoral dissertation, Stanford University, 1978.

Chicago Board of Trade, *Annual Reports*, Chicago, 1880–1886.

Bureau of Mines, *Minerals Yearbook*, Washington, various years.

An Analysis of the Selection of Arbitrators

By David E. Bloom and Christopher L. Cavanagh*

This paper analyzes data on union and employer rankings of different panels of arbitrators in an actual arbitration system. A random utility model of bargainer preferences is developed and estimated. The estimates indicate that unions and employers have similar preferences: in favor of lawyers, more experienced arbitrators, and arbitrators who seem to have previously favored their side. Tests of whether bargainers rank arbitrators strategically reveal no evidence of such behavior.

Arbitration is a rapidly growing method for resolving disputes. It is used widely in the United States to resolve private disputes arising under collective bargaining agreements and commercial contracts, to resolve certain types of civil disputes, and to set wages in the public sector. Although arbitration has been applied in a wide range of settings and takes numerous forms, the central feature of virtually all arbitration mechanisms is that they involve a third party, that is, an arbitrator or panel of arbitrators, hearing and deciding how a dispute is to be resolved. In this respect, arbitration may be viewed as a private-sector analog to the court system with arbitrators performing similar functions to judges. Arbitration tends, however, to be a cheaper, quicker, and less formal method of dispute resolution than the court system. Arbitration systems also tend to provide disputing parties with greater latitude in choosing "judges" (i.e., arbitrators) than do court systems.[1]

Among persons who are regularly involved in arbitration, it is generally accepted that good arbitrators are the key to "good arbitration." However, there seems to be far less consensus about the meaning of the phrase "good arbitrators." For example, is a good arbitrator someone who favors your side, or someone who is painstakingly fair in arriving at a decision? Is it someone who has extensive experience with the type of dispute at hand, or someone with good common sense and the ability to analyze, interpret, and judge? Is it someone who tries to appease the parties by splitting decisions, or someone who strictly "calls them as he or she sees them?"

Our purpose in this paper is to address these and other related questions by presenting an empirical analysis of the selection of arbitrators. We do this by analyzing a remarkable set of data on the preferences of unions and employers for different arbitrators under New Jersey's Fire and Police Arbitration Law. According to this law, disputes over terms and conditions of employment involving New Jersey's organized public safety employees and the governments that employ them must be resolved by arbitration. Arbitrators, whose awards are binding by law, are chosen by the bargainers from a roster of roughly seventy names maintained by the New Jersey Public Employment Relations Commission (PERC). In cases in which the bargainers are unable to negotiate an agreement, PERC circulates a list of seven arbitrators and their resumes to the parties, each of which is instructed to veto three names and to rank, in order of their preferences, the remaining four names. PERC then appoints as arbitrator the indi-

*Department of Economics, Harvard University, Cambridge, MA 02138. We thank Orley Ashenfelter, two referees, and seminar participants at Harvard University, Princeton University, the NBER, and MIT for helpful comments. This research was supported by NSF grant no. SES-8309148.

[1] For some additional comparisons of arbitration and the court system, in which the advantages of arbitration are stressed, see the text of Chief Justice Warren Burger's (1985) remarks to the American Arbitration Association and the Minnesota State Bar Association.

vidual who was not vetoed by either side, and whose combined rank is highest; rank ties are broken randomly by PERC.

Using information on employer and union rankings of different panels of arbitrators, along with information on the characteristics of the arbitrators, we attempt to provide direct evidence on the following three issues:

1) How similar are the preferences of unions and employers with respect to a given panel of arbitrators?

2) What characteristics of arbitrators do unions and employers find desirable or undesirable? Do the unions and employers attach the same or different weights to specific characteristics? and

3) Do unions or employers engage in strategic behavior in ranking arbitrators?

In proceeding this way, we also hope to shed light on three broader issues. First, there has recently developed in the academic literature a body of theoretical work on the subject of bargaining and arbitration.[2] The basic premise of most of this work is that arbitration is simply a mechanism for distributing income between conflicting interests. In contrast, institutional economists and labor practitioners place greater emphasis on arbitration as a mechanism for helping disputants identify and reach efficient outcomes. In their view, arbitrators are professional gatherers and processors of information who play a highly constructive role in a bargaining process which is better treated as a cooperative attempt at problem solving than as direct economic conflict. By analyzing the similarity of union and employer preferences for individual arbitrators and for different arbitrator characteristics, we suspect that much can be learned about the general issue of whether collective bargaining is primarily an institution of cooperation or conflict.

Second, one of the most important characteristics of arbitration systems is that they may be designed in different ways. Indeed, one key dimension in which arbitration systems differ involves the mechanism for selecting the arbitrator. Most mechanisms take account of the parties' preferences, either through a rank-four/veto-three system like New Jersey's, or by allowing each party to successively veto a name from an odd-numbered list of three or more arbitrators. Other arbitration systems appoint arbitrators on a purely rotating basis from a list agreed to in advance by potential disputants or established by a third party such as the state. A final system involves the appointment of a single individual or panel of individuals to arbitrate all disputes involving a particular set of parties and arising in a specified period of time.

The key feature of all of these systems is that they guarantee the appointment of an arbitrator without requiring explicit agreement (or even face-to-face contact) by two parties who are unable to reach agreement on some other (substantive) matter. In addition, these systems all prescreen individuals before they are added to the master list of eligible arbitrators. However, the first two systems provide for an additional level of screening by the parties prior to the appointment of an arbitrator to hear a particular case. This additional level of screening is said to contribute to the legitimacy of the arbitrator and his award in the eyes of the parties. However, it can also contribute to delays in the arbitration process, which is one of the most frequently cited complaints about arbitration. Thus, it is interesting to ask whether the appointment of arbitrators can be left entirely in the hands of the state or some impartial organization like the American Arbitration Association, or whether it is important to take account of the parties' preferences on a case-by-case basis. We will address this question below when we analyze the strength of union and employer preferences for different members of a set of "prescreened" arbitrators.[3]

[2] See, especially, the work of Vincent Crawford (1979, 1982) and Henry Farber (1980) and Farber and Harry Katz (1979). For a useful review of selected empirical analyses of some of the theoretical issues raised in these papers, see Orley Ashenfelter (1985).

[3] In many respects, the selection of an arbitrator is just one example of a general class of social choice problems in which two or more economic agents must collectively decide an intermediate or final outcome of some economic game. Voting for public officials,

Third, our study raises important questions about the possibility of stategic behavior and its treatment in empirical analysis. In particular, it is well known that the outcomes of voting mechanisms can often be manipulated by the strategic misrepresentation of preferences. Although it is natural to address this problem by directly estimating a structural model of the underlying economic game, the complexity and dimensionality of the game may render this approach infeasible, as it does in our case. Thus, we develop some indirect tests that we think will let us "back our way" into strategic behavior if it is there. We suspect that the type of indirect approach we propose may be useful in a variety of game-theoretic settings in which the games are too complicated to solve.

In Section I, we provide some institutional background on the selection of arbitrators and describe our data. In Section II we set out a simple random utility model that we use to represent the preferences of employers and unions for arbitrators with different characteristics. We also present the likelihood function we maximize to estimate the parameters of this model.[4] In Section III we present a descriptive summary of the data. We also present estimates of the econometric model as well as two alternative models we estimate to test for the presence of strategic behavior. Section IV summarizes and concludes the paper.

I. Institutional Background

Most of the data analyzed in this paper were drawn from PERC's arbitration records. First, we collected information on the lists of arbitrators sent by PERC to disputing parties along with the preference rankings returned to PERC by the parties. We focused only on cases involving 1980 contract negotiations. That was the third year of operation of the New Jersey arbitration system and the third (and final) year in which PERC used its original master list of eligible arbitrators to form panels.[5] Thus, we felt that by 1980 the parties had reasonably good information about the arbitrators on which to base their preference rankings. In many cases, at least one of the parties did not strictly follow PERC's request for a preference ranking. Sometimes parties ranked more than four names on the list; other times the parties vetoed more than three names; in a few cases a party responded to PERC by saying that all seven names were equally acceptable; there were also a number of cases in which a party either failed to express its preferences to PERC or its preferences were simply not recorded in the PERC records.[6] Altogether we collected information on 193 arbitration panels. Of these, 75 are perfect in the sense that *both* parties ranked four names and vetoed three names. It should also be noted that many (indeed, most) of the cases for which PERC circulated arbitration panels did not end up being arbitrated. In other words, many disputants were able to reach voluntary settlements after the arbitration

reaching committee decisions, choosing reals estate appraisers (for example, in cases of eminent domain), and determining the recipients of different honors and awards are all examples. But the examples which are most closely akin to the problem of selecting an arbitrator are the problems of judge and jury selection. In the case of judges, prescreening is substantial (for example, all federal judges must be nominated by the president and confirmed by the Senate), although assignments are random except for the practice of "forum shopping" and recusing in situations where there are conflicts of interest. On the other hand, prescreening is minimal in the process of jury selection since federal and state jury selection laws generally require that potential jurors be "selected at random from a fair cross section of the community" (P.L. 90–274, 82 Stat. 53). However, jury selection procedures offer opportunities to remove jurors both for cause and, although to a lesser extent, without cause. These procedures, known as *voir dire*, are modeled in Arthur Roth et al. (1977).

[4] Detailed derivations of all of the likelihood functions we use and their properties are presented in the Appendix to our 1985 paper, which is available on request.

[5] PERC revises its master list of eligible arbitrators every three years.

[6] This unbalanced data configuration complicates the descriptive presentation of the data below. However, the econometric model developed in Section II is ideally suited to this type of problem and makes efficient use of *all* available information.

panel was circulated but before a binding arbitration award was rendered. This characteristic of the bargaining/arbitration process in New Jersey explains why some parties did not report their arbitrator preferences to PERC (i.e., either their case was settled before the due date for reporting their preferences, or they expected that it would not end up in arbitration).

Second, we collected information on the characteristics of the sixty-nine arbitrators on PERC's master list. This information was derived from a variety of sources including PERC's 1978 and 1979 interest arbitration records and awards, PERC's 1979 grievance arbitration records, and the arbitrators' resumes. In collecting this background information, we were guided by the institutional literature on the relevant characteristics of arbitrators and by conversations with labor relations practitioners.[7] Roughly speaking, the relevant characteristics of arbitrators seem to fall into four categories: 1) impartiality; 2) consistency; 3) training; and 4) experience.[8]

Impartiality refers to an arbitrator's lack of predisposition to rule in favor of one side or the other. This characteristic is usually judged by considering an arbitrator's prior decisions. It appears to be the most important characteristic of an arbitrator since no party is likely to be satisfied with an arbitrator it perceives to be biased against its position. There has even been some debate over whether disputants prefer arbitrators who they perceive to be biased in their favor. On the one hand, such bias is desirable because it suggest a higher probability of receiving a favorable arbitration decision. But, on the other hand, it damages the integrity of the institution of arbitration and does not promote the legitimacy and mutual acceptability of the arbitrator's award.

Consistency refers to the extent to which an arbitrator decides cases solely on their merits, without reference to his (or her) "box score" of previous decisions. The parties' concern with the consistency of arbitrators stems from their awareness that many arbitrators derive considerable income from arbitrating, and therefore may have an incentive to "split their awards" in order to appear impartial and maintain their acceptability. Such a practice greatly threatens the institution of arbitration, since it suggests that a certain fraction of cases will be won by a given party, not on their merits, but simply because they are brought before an arbitrator. Consistency is most often judged by subjectively reviewing an arbitrator's previous awards to see that similar decisions were reached in similar cases.

A third dimension along which arbitrators differ is their training. Most labor arbitrators are lawyers, undoubtedly because legal training is well-suited to analyzing and judging the vast majority of labor disputes, that is, disputes over the terms of existing contracts, also known as grievances. However, the New Jersey system involves disputes over the terms of new contracts, that is, disputes of interest, with the most common and important issue in dispute being wages. Thus, one might expect that training in other areas, and especially in economics, might be particularly desirable to the parties.

Arbitration experience is another important characteristic of prospective arbitrators. Practitioners generally regard this characteristic as a measure of an individual's expertise as an arbitrator and usually insist upon an experienced arbitrator in cases involving complex or otherwise difficult issues.

[7]We relied particularly on material contained in Lois MacDonald (1948) and Frank Elkouri and Edna Elkouri (1985), and on conversations with Ben Fischer (former head of the Arbitration Department of the United Steelworkers of America) and Richard Reilly (Regional Director of the American Arbitration Association, Boston Region).

[8]The per diem that arbitrators charge is also a way in which they differ. Most *ad hoc* arbitrators, for example, presently charge between $200 and $600 per day, plus expenses. Although little is known about the extent to which cost influences the selection of an arbitrator, it does drain the parties' funds and has been argued to be an important determinant of the use of arbitration in the case of financially small disputants (see Bloom, 1981). In the New Jersey system under study, PERC establishes a maximum per diem rate which is the rate charged by nearly all of the arbitrators on the master list. As a result, there is very little variation in arbitration fees across arbitrators. Thus, we do not include this variable in our empirical analysis.

At a more theoretical level, it seems likely that experience is desirable because it reduces the uncertainty that risk-averse disputants have over the outcome of arbitration and thereby reduces its (indirect) cost.[9] Indeed, the acquisition of experience is generally regarded to be the most significant hurdle faced by aspiring labor arbitrators in the United States.[10]

II. A Model of Arbitrator Selection

In this section we outline a simple model of arbitrator selection. We start by considering a situation in which an employer (E) and a union (U) are unable to reach agreement on a vector of contract items, although they do agree (or are compelled by law) to have their dispute resolved by arbitration. A list of seven potential arbitrators along with information on their qualifications is circulated among the two parties by an impartial organization. Each party is instructed to veto three names and rank the remaining four names in order of their preference. The individual who is not vetoed by either side and who has the lowest combined rank is appointed to hear the case; rank ties are broken by coin tosses.[11]

We now assume that each party has a preference ordering defined over the entire set of possible arbitration outcomes and that these preference orderings can be represented by well-behaved von Neumann-Morgenstern utility functions. We also assume that each party has one set of prior beliefs about the distribution of arbitration outcomes for each of the seven potential arbitrators. Finally, we assume that these sets of prior beliefs depend on the arbitrators' qualifications, some of which are observed.

In this model, an arbitrator's characteristics determine a party's prior beliefs about the arbitrator's decision. These beliefs, in turn, determine the expected utility that a party associates with that arbitrator. It follows that there exist direct mappings of arbitrator characteristics into an expected utility for each party. For each potential arbitrator (i), we write these expected utilities as a linear function of the arbitrator's characteristics.

$$Y_{Ei} = X_i\beta_E + \varepsilon_{Ei} \qquad (i=1,2,\ldots,7) \tag{1a}$$

$$Y_{Ui} = X_i\beta_U + \varepsilon_{Ui} \qquad (i=1,2,\ldots,7) \tag{1b}$$

where Y_{Ei} and Y_{Ui} are the expected utilities the employer and union associate with arbitrator i; X_i is a vector of observed characteristics of arbitrator i; β_E and β_U are vectors of unknown (reduced-form) parameters characterizing the preferences of the employer and union for different arbitrator characteristics; and ε_{Ei} and ε_{Ui} represent random utility effects.

Given this framework, the simplest model one could adopt for arbitrator selection would treat each party as ranking arbitrators in order of their expected utilities. In other words, each party's stated ranking is assumed to coincide with the ordering of its true preferences. More complicated models which account for each party's incentives to rank arbitrators some other way are possible, too. For now, however, we shall work with the simple model and defer consideration of alternative models to Section III, Part C.

In order to make this model of arbitrator selection empirically tractable, it is necessary to treat the unobserved characteristics ε_{Ei} and ε_{Ui} ($i=1,\ldots,7$) as random variables. We do this by assuming that these random variables have independent extreme value distri-

[9]This conclusion requires the assumption that each party is equally risk averse if negotiated outcomes are a possibility.

[10]Recently, however, the high cost, long delays, and general shortage of experienced arbitrators have caused some individuals to question the importance of experience. For example, a system known as expedited arbitration was adopted by labor and management in the basic steel industry in 1971. Under this system, unresolved employee grievances that do not require precedent-setting rulings are arbitrated by a rotating panel of young, inexperienced arbitrators (mostly lawyers) who decide the case for a relatively small fee within two weeks of the decision to arbitrate. Although there has yet to be an in-depth study of this system, its growing utilization in the basic steel industry, in other United Steelworker contracts, in some United Mineworker contracts, and in the U.S. Postal Service provides some evidence of its success.

[11]In this section we ignore the possibility that a voluntary settlement can occur after the appointment of an arbitrator but prior to arbitration.

butions:

$$(2)\ Pr(\varepsilon_{pi} \leq x) = \exp(-e^{-x}) \quad (p = E, U) \quad (i = 1, \ldots, 7).$$

This distributional assumption is quite common in random utility models of this general form.[12] It is also quite convenient, both analytically and computationally, in the present application.

We derive the likelihood function for each party's rank choices by developing an expression for the probability of a particular ranking. For example, suppose the union ranked a list of seven arbitrators in the following order: (1,2,3,4, *Veto*, *Veto*, *Veto*). The likelihood of this ranking is simply the probability that $(Y_{U1} > Y_{U2} > Y_{U3} > Y_{U4} > Y_{U5}, Y_{U6}, Y_{U7})$. The general formula for the probability of a particular ranking is given by

$$(3)\ \prod_{j=1}^{K} MNL(A(j) \mid \text{Remaining choices}, \beta)$$

where $A(j)$ denotes the arbitrator who receives the jth rank, K is the number of arbitrators that receive a rank (for example, four, in the standard case), and MNL is the multinomial logit probability that a party with preference parameters β will most prefer arbitrator $A(j)$ given the option of choosing from the $8-j$ least-preferred arbitrators. This probability is defined as

$$(4)\quad MNL(A(j) \mid \text{Remaining choices}, \beta) = \exp(X_{A(j)}\beta) \Big/ \sum_{i \in \phi_{8-j}} \exp(X_{A(i)}\beta)$$

where ϕ_{8-j} refers to the remaining choice set of $8-j$ least-preferred arbitrators.

The probability in equation (3) is simply the product of the multinomial logit probabilities of ranking an arbitrator from among those arbitrators that have not already been ranked. Notice that this model is extremely amenable to dealing with the type of unbalanced data configuration we face. For example, in cases where more than four arbitrators are ranked, the likelihood of the ranking will simply consist of the product of more than four multinomial logit probabilities. Thus, this model makes efficient use of all of the preference ordering information we have available. It should also be noted that this likelihood function is globally concave. This property guarantees a unique maximum if the model is asymptotically identified and typically assures numerical stability in computing maximum likelihood estimates.

III. Empirical Results

A. *Descriptive Statistics*

We begin our discussion of empirical results by presenting a descriptive summary of the data. Table 1 presents the distribution, across arbitration panels, of the number of arbitrators not vetoed by either side. If employer and union preferences tend to be in direct conflict, one would expect most panels to yield only a single jointly ranked arbitrator. Alternatively, if employer and union preferences tend to be similar, one would expect most panels to yield four jointly ranked names. As the first row of Table 1 makes clear, neither of these extremes appears to be true. Over 80 percent of the panels yielded two or three jointly acceptable arbitrators with an average overlap of nearly 2.5 arbitrators per panel. Observe also the last row of Table 1 that presents the distribution of the number of jointly ranked names under the assumption that the parties' rankings of arbitrators are independent of each other. A *chi*-square test comparing the observed distribution (row A) to the independent-rankings distribution (row D) yields a test statistic of 11.2, which is statistically significant at the .05 level (3 degrees of freedom). This result suggests that the parties' rankings are not independent. Moreover, the biggest contribution to the test statistic comes from the proportion of lists with four overlapping names, providing some evidence of positive correlation in the parties rankings. In addition, it is worth noting that the distribution of jointly ranked names is not statistically significantly different in the subset of

[12] See, for example, Daniel McFadden (1982).

TABLE 1—PERCENT DISTRIBUTION OF NUMBER OF ARBITRATORS NOT VETOED BY EITHER PARTY[a]

	1	2	3	4	Average Overlap
A. All Cases ($N=75$)	9.3	44.0	37.4	9.3	2.47
B. Arbitrated Cases ($N=29$)	6.9	41.4	37.9	13.8	2.59
C. Negotiated Cases ($N=46$)	10.9	45.7	37.0	6.5	2.39
D. Independent Rankings	11.4	51.4	34.2	2.9	2.28

[a]These statistics, as well as those in Table 2 and Figure 1, are based on the 75 cases in which both sides ranked four arbitrators and vetoed three.

29 cases that ended up being arbitrated than in the subset of 46 cases that ended up being negotiated (i.e., a *chi*-square test for the equality of the true distributions yields a test statistic of 1.4, which is not in the 5 percent tail of a χ^2 distribution with 3 degrees of freedom).

Since Table 1 is not informative about the closeness of employer and union rankings within each panel of arbitrators, we have computed a rank correlation coefficient (ρ) for each of the 75 "perfect" cases:

$$\rho=\frac{1}{28}\sum_{i=1}^{7}R_{Ei}R_{Ui}-4, \tag{5}$$

where R_{Ei} and R_{Ui} are the ranks assigned by the employer and the union, respectively, to the ith arbitrator on the panel.[13] In Figure 1 we plot the (observed) frequency distribution of this statistic. As can be seen from the plot, there are some cases in which the parties' rankings are very different (i.e., $\rho\leq-.5$), some cases in which they are very similar (i.e., $\rho\geq.5$), and many cases in which they seem to be uncorrelated. However, on balance, the plot seems to produce evidence of a slight tendency for the employer and union rankings to be positively correlated ($\bar{\rho}=.13$).[14]

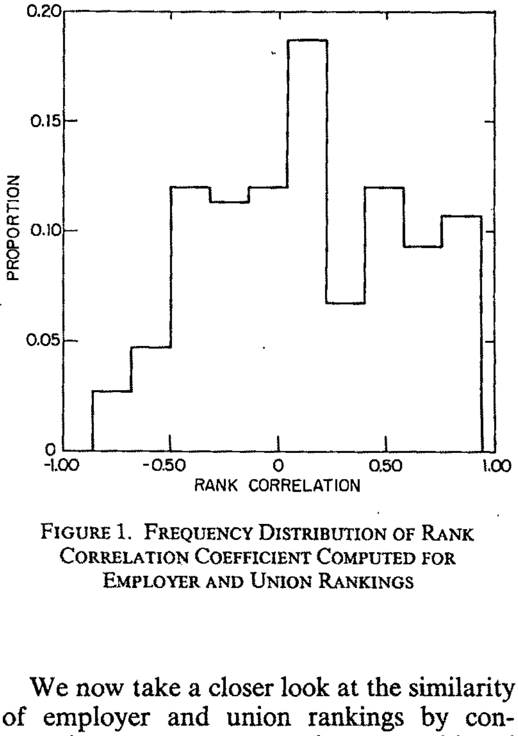


FIGURE 1. FREQUENCY DISTRIBUTION OF RANK CORRELATION COEFFICIENT COMPUTED FOR EMPLOYER AND UNION RANKINGS

We now take a closer look at the similarity of employer and union rankings by constructing a two-way contingency table of these rankings and formally testing the hypothesis that they are independent. This contingency table, which reports the number of times that arbitrators listed on the 75 panels received each of twenty-five possible combinations of employer and union ranks, is presented in Table 2. This table also presents (in parentheses) theoretical frequencies for each cell computed under the hypothesis that the rankings are independent.[15] The hypothesis

[13]If all arbitrators on the panel were assigned a rank between 1 and 7 by each party, ρ would always lie in the interval $[-1.0, 1.0]$. However, because only four arbitrators of the seven listed on each panel receive a rank, we were forced to assign a rank of 6.0 to the three vetoed arbitrators (i.e., the median of 5, 6, and 7). As a result, our estimates of ρ must lie in the interval $[-.86,.93]$.

[14]We have not worked out the distribution of ρ for the incomplete rankings case. Thus, we are unable to construct a formal test.

[15]For example, the expected number of arbitrators vetoed by the union *and* ranked first by the employer is 32.1 ($=3/7\cdot1/7\cdot525$).

TABLE 2—TWO-WAY DISTRIBUTION OF OBSERVED EMPLOYER AND UNION RANKINGS[a]

Employer Rank	Union Rank 1	2	3	4	Veto	Total
1	21 (10.7)	10 (10.7)	11 (10.7)	5 (10.7)	28 (32.1)	75
2	8 (10.7)	16 (10.7)	15 (10.7)	14 (10.7)	22 (32.1)	75
3	7 (10.7)	12 (10.7)	14 (10.7)	7 (10.7)	35 (32.1)	75
4	10 (10.7)	10 (10.7)	10 (10.7)	15 (10.7)	30 (32.1)	75
Veto	29 (32.1)	27 (32.1)	25 (32.1)	34 (32.1)	110 (96.4)	225
Total	75	75	75	75	225	525

[a] Theoretical frequencies are reported in parentheses.

of independence is tested using the familiar χ^2 statistic which has the value 36.4 for this table. Since this statistic has 15 degrees of freedom, we reject the hypothesis of independence at all conventional significance levels (for example, the critical value of a χ^2_{15} random variable at the 99 percent level is 32.8).

Before leaving Table 2, it is interesting to note that the (1,1) cell makes the biggest contribution to the χ^2 statistic. In other words, the hypothesis of independence is rejected largely because the unions and the employers ranked the same individual first in 21 of the 75 panels they reviewed. Indeed, the observed frequencies are greater than the theoretical frequencies for all of the diagonal cells in Table 2. This provides some further evidence that employer and union preferences tend to be at least moderately similar.

We now describe the characteristics of the New Jersey arbitrators. To begin with, 45 percent of the arbitrators have law degrees and 12 percent are Ph.D. economists; the remaining 43 percent are labor relations practitioners, some of whom have Ph.D.s in labor relations or other areas. All of the arbitrators, of which there are only four women, have mediation or fact-finding experience in public-sector wage disputes. In addition, two-thirds of the arbitrators had rendered at least one interest arbitration award in the New Jersey system during its first two years of operation, although only 12 percent of the arbitrators had rendered five or more awards. On average, the New Jersey arbitrators have about fifteen years of arbitration experience, with the range of experience being from four to forty years. Nearly two-thirds of the arbitrators had been appointed to hear two or more grievance arbitration cases in the New Jersey public sector in 1979. About one-half of the arbitrators are members of the National Academy of Arbitrators.

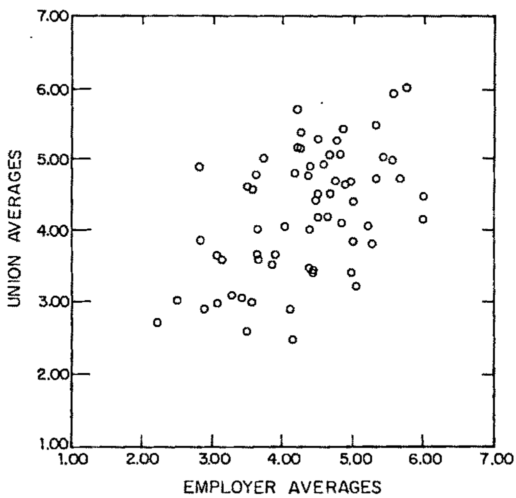


FIGURE 2. AVERAGE RANKINGS OF INDIVIDUAL ARBITRATORS, BY EMPLOYERS AND UNIONS

In Figure 2 we plot the average rankings received from the employers and the unions by the sixty-five arbitrators who received at least two rankings from each party.[16] In

[16] Different arbitrators appeared on different numbers of panels because PERC generated the panels randomly

computing these averages we were able to use ranking information from most of the 193 panels by assigning vetoed arbitrators a rank equal to the median of the unassigned ranks. As Figure 2 makes clear, there is considerable dispersion across arbitrators in the average rankings they received from the parties. This provides further evidence that the parties prefer some arbitrators to others. Moreover, the average rankings in Figure 2 exhibit a fairly high positive correlation ($\rho = .51$), providing a further indication of similarity between the parties' preferences. Of course, one weakness of the average-rank statistics plotted in Figure 2 is that they do not control for the characteristics of the other arbitrators who appeared on the same lists. However, controlling for those characteristics is accomplished by our structural model of arbitrator selection, to which we now turn.

B. *Estimation Results*

Table 3 presents estimates of the employer and union preference parameters of the random utility model in equation (1). Positive signed coefficients indicate that a particular characteristic tends to increase the expected utility of an arbitrator and, therefore, the likelihood that the arbitrator receives a favorable rank. Negative coefficients indicate the opposite. The estimates were obtained by maximizing the log of the likelihood function implied by equation (3) for a particular specification of arbitrator characteristics. The maximization was accomplished on a personal computer using the modified quadratic hill-climbing method (*GRADX*) proposed by Stephen Goldfeld and Richard Quandt (1972). All of the rankings data available from the 193 arbitration panels were used in the estimation (i.e., we used 129 employer rankings and 160 union rankings). Since the estimates are maximum likelihood, standard tests of their significance can be performed based on their asymptotic normality.

In specifying the vector of arbitrator characteristics, we paid close attention to the institutional literature on arbitrator selection reviewed in Section I. Thus, we attempted to capture the training dimension of an arbitrator's characteristics by including dummy variables for lawyers and economists, with all other arbitrators comprising the reference category. We have no strong priors on the effect of these variables, although it does seem likely that economists would be viewed as best able to resolve the wage and benefit disputes that are central to most negotiations in New Jersey. Our specification also includes variables reflecting an arbitrator's experience. Specifically, we include the number of grievance arbitration appointments each arbitrator received in 1979, and the number of conventional and final-offer arbitration awards each arbitrator rendered in New Jersey prior to 1980.[17] If experience is truly a desirable characteristic, we would expect both of these measures, which are derived from independent arbitration systems, to be positively associated with each parties' preferences for an arbitrator.

Finally, we attempt to control for an arbitrator's impartiality in several ways. First, we include a variable (*FAVU*) defined to be the difference between the number of final-offer cases decided in favor of unions and the number decided in favor of employers in the years 1978 and 1979. This variable is nonzero for about 55 percent of the arbitrators and takes on values between -4 and 4. If this variable adequately measures the tendency of an arbitrator to be more sympathetic to one side than the other, and if the

---

and because a number of arbitrators requested that their names not be circulated actively throughout the entire year.

[17]Final-offer arbitration is utilized to resolve about three-fourths of the bargaining disputes arising under the New Jersey statute. Under final-offer arbitration, the arbitrator is constrained to render an award which consists of one or the other of the bargainers' final positions. Most of the remainder of the New Jersey cases are resolved by conventional arbitration in which the arbitrator renders a decision which consists of his or her best judgment of a fair settlement and which may be a compromise between the parties' final offers. For a more detailed description of the New Jersey statute, see Bloom (1980). For an analysis of arbitrator decision making under the different forms of arbitration in New Jersey, see Ashenfelter and Bloom (1983, 1984).

TABLE 3—MAXIMUM LIKELIHOOD ESTIMATES OF EMPLOYER AND UNION PREFERENCE FUNCTIONS, COMPLETE RANKINGS (*CR*) MODEL[a]

Variable	Definition	Sample Average for 69 Arbitrators	Employer Preferences (1)	(2)	(3)	(4)	Union Preferences (1)	(2)	(3)	(4)
LAWYER	Arbitrator has a law degree	.45	.204	.177	.174	.101	.158	.109	.194	.219
			(.114)	(.121)	(.113)	(.112)	(.100)	(.105)	(.099)	(.098)
ECONOMIST	Arbitrator is a Ph.D. economist	.12	.478	.437	.433	.410	−.782	−.865	−.746	−.744
			(.205)	(.214)	(.207)	(.208)	(.219)	(.225)	(.220)	(.223)
GRAPTS79	Number of grievance arbitration appointments in NJ public sector in 1979	3.99	.076	.076	.076	.070	.078	.078	.080	.074
		(4.66)	(.013)	(.013)	(.013)	(.015)	(.012)	(.012)	(.012)	(.014)
NFOA	Number of final-offer arbitration (*FOA*) awards rendered in NJ in 1978–79	1.73	.076	.060	.058		.005	−.025	.026	
		(2.56)	(.022)	(.032)	(.020)		(.019)	(.028)	(.018)	
NTOT	Total number of conventional and final-offer awards rendered in NJ in 1978–79	2.23				.046				.035
		(3.12)				(.021)				(.018)
FAVU	Number of *FOA* awards decided in favor of unions minus number of *FOA* awards decided in favor of employers, 1978–79	.22	−.077	−.085			.097	.079		
		(1.39)	(.032)	(.034)			(.030)	(.032)		
FAVUHAT	Predicted value of *FAVU* (See text)	.48		.076				.163		
		(.76)		(.116)				(.105)		
FAVDIFF	*FAVU* minus *FAVUHAT*	−.27			−.085				.082	
		(1.17)			(.034)				(.032)	
AVEDEV	Arbitrator's average percent wage increase, expressed as deviation from grand average of wage increases	−.05				−.046				.108
		(.88)				(.054)				(.049)
Log likelihood			−769.1	−768.9	−768.9	−772.8	−931.0	−929.7	−932.8	−933.6
($\beta = 0$)				(−827.4)				(−1018.0)		

[a] Estimated standard errors are reported in parentheses.

parties prefer arbitrators they perceive to be relatively sympathetic to their position, this variable should be associated with more favorable rankings from the union and less favorable rankings from the employer. Second, we construct a variable (*FAVUHAT*) which measures the *expected difference* between the number of final-offer cases decided in favor of unions and employers. This measure is constructed using estimates of the unconstrained arbitrator decision-making functions reported in Orley Ashenfelter and

Bloom (1984, Table 2, col. 1 for 1978 and 1979) along with information on the final offers in the 1978 and 1979 cases. The inclusion of this variable in the empirical model refines the first measure of impartiality by controlling for an important subset of facts in the final-offer cases. Thus, under the assumption that the parties evaluate the record of each arbitrator's final-offer decisions in light of the quality of the final offers they were forced to choose between, our expectation is that the coefficient of *FAVUHAT* will be opposite in sign but have the same absolute magnitude as the coefficient of *FAVU*. Third, as a simple alternative measure of arbitrator bias, we include a variable (*AVEDEV*) defined as the average wage increase awarded by each arbitrator in their 1978 and 1979 cases. This variable, which incorporates information from both conventional and final-offer cases, is expressed as a deviation from the average of all awards rendered in 1978–79; it has the value zero for arbitrators who made no awards. If case facts are similar across bargaining disputes, this variable will provide a reasonable measure of each arbitrator's tendency to favor the union or the employer.

The estimates in Table 3 reveal a number of interesting features about employer and union preferences for arbitrators. First, both the employers and the unions tend to prefer individuals with law degrees to labor relations practitioners. However, employers prefer economists to both of these groups, whereas unions prefer both of these groups to economists. Perhaps this somewhat surprising result is explained by the fact that economists are likely to be heavily influenced by efficiency considerations, whereas lawyers are more likely to place greater emphasis on equity.

Second, the three measures of an arbitrator's experience (*GRAPTS79*, *NFOA*, and *NTOT*) typically have positive coefficients in both the employer and union equations. This indicates that employers and unions both prefer more experienced arbitrators to less experienced arbitrators, controlling for variables reflecting their training and impartiality. It is, however, possible to interpret the experience variables as controlling for arbitrator-specific characteristics which are not in the model (akin to fixed effects) since, to some extent, they reflect arbitrator's past popularity. Since it is not possible for us to distinguish empirically between these alternative interpretations of the experience variable, we conclude simply that our results are not inconsistent with the hypothesis that experience is a desirable characteristic, as emphasized in the institutional literature on arbitrator selection.

Third, our measures of arbitrator bias, that is, their past tendency to favor unions, have negative coefficients in the employer equation and positive coefficients in the union equation, whether or not we control for the case facts. This indicates that employers tend to give poorer rankings to arbitrators who seem to have favored unions in the past, whereas unions tend to prefer such arbitrators. This finding is consistent with our expectation that the parties do not like arbitrators whom they perceive to be biased against them. Note also that the coefficients of *FAVUHAT* are not significant at the 5 percent level in either the employer or the union equations, and that their inclusion in the equations does not substantially change the magnitudes of the coefficients of *FAVU*. In addition, in the employer equation the coefficient of *FAVUHAT* is opposite in sign and of roughly the same magnitude as the coefficient of *FAVU*. However, in the union equation the two coefficients have the same sign and, counter to our expectation, we reject the hypothesis that they sum to zero at the 5 percent level. Nonetheless, if we constrain the coefficients of *FAVU* and *FAVUHAT* to be equal and opposite in sign (i.e., by entering *FAVDIFF*), the measure of bias has the theoretically expected effect and is close in magnitude to the bias coefficient in the other models.

The last point worth noting about the results in Table 3 relates to the magnitudes of the coefficients. These are difficult to interpret since their scale is determined by our assumption that the ε_{Ei} and ε_{Ui} are drawn independently from extreme value distributions as defined in equation (2). However, one may gauge the magnitude of the coefficients by measuring their values in relation

to the standard deviation of the ε's. The latter are fixed by our distributional assumption and equal $\pi/\sqrt{6}$ (≈ 1.28). Viewed in this way, we see that all of the coefficient estimates are relatively small in magnitude. For example, a one-standard-deviation change in the number of grievance arbitration appointments only represents a change of about one-fourth of a standard deviation of ε. Thus, while our estimates of the effect of arbitrator's characteristics on the parties' rankings are reasonably precise, they are also quite small.[18] Given the similarity of results from alternative specifications (reported in our 1985 paper) that allow for nonlinearities and that include alternative measures of experience, we suspect that the large amount of noise we are observing indicates that the parties are relatively indifferent to many of the arbitrators in the New Jersey system.

C. *Testing for Strategic Behavior*

As noted earlier, a critical assumption of our analysis is that the observed rankings data reveal the true preferences of the parties. This would be true if each party perceived that it had no incentive to misstate its true preferences. However, it is not difficult to imagine a scenario in which the parties do have incentives to rank arbitrators strategically. For example, suppose the union's true preference ranking of a list of seven arbitrators is [1, 2, 3, 4, *Veto*, *Veto*, *Veto*] and the employer's true preference ranking is [1, 4, 3, 2, *Veto*, *Veto*, *Veto*]. Although not identical, these preference orderings do have important similarities: they both rank the first arbitrator as most preferred and they both veto the fifth, sixth, and seventh arbitrators. However, if either party has some information about the other party's preferences, it is quite likely that it will have an incentive to misstate its true preferences. For example, if each party expects the other to veto the last three arbitrators on the list, but has no idea how the other party will rank the first four arbitrators, they will both try to ensure the appointment of the first arbitrator by presenting identical rankings of [1, *Veto*, *Veto*, *Veto*] for the first four arbitrators. Thus, the two parties' revealed preferences might appear quite similar even though the only correspondence between the true and revealed preferences of each party would be the assignment of the most favorable rank to the first arbitrator. Other examples, in which similar preferences can lead to dissimilar rankings, are possible, too, This type of behavior is particularly distressing since it can lower both parties' welfare (see our 1986 paper). Moreover, it suggests that the estimates presented so far in this section may not reflect the parties' true preferences.

A natural way to address this issue is to model the ranking of arbitrators as a game in which each party must choose a strategy, that is, a ranking, given its subjective opinion about the ranking of the other party. A natural definition of equilibrium strategies for this game would be the Bayesian Nash equilibrium which has the characteristic that neither party can increase its expected utility by unilaterally changing its strategy. In principle, one could solve for each party's Nash strategy and derive a likelihood function that could be maximized to estimate each party's preference parameters conditional on this strategic behavior. However, in practice, the dimensionality of the problem makes this infeasible unless substantial structure is placed on the problem, which we prefer not to do.[19] Instead, we will develop an empirical approach to this issue.

Our test for strategic behavior primarily involves the estimation of alternative random utility models that use different subsets of information available in the arbitrator rankings data. Under the null hypothesis that the data reveal the parties' true preferences, estimates of the parameters of these models

[18] Nevertheless, the arbitrator characteristics included in the model do have significant explanatory power when considered jointly, as judged by the difference between the maximized log likelihood and the log likelihood evaluated at $\beta = 0$.

[19] In principle, each side has 840 distinct pure ranking strategies, of which at least several hundred are undominated. Clearly, solving for the Nash equilibrium strategies in such a game is computationally infeasible (see our 1986 paper for a description of equilibria in these and other related games).

TABLE 4—MAXIMUM LIKELIHOOD ESTIMATES OF EMPLOYER AND UNION PREFERENCE FUNCTIONS, MULTINOMIAL LOGIT MODEL (*MNL*) AND RANK/VETO MODEL (*RV*)[a]

	Employer Preferences		Union Preferences	
Variable	*MNL*	*RV*	*MNL*	*RV*
LAWYER	.386	.146	.088	.169
	(.226)	(.119)	(.193)	(.108)
ECONOMIST	.669	.355	−.574	−.814
	(.391)	(.218)	(.452)	(.229)
GRAPTS79	.103	.082	.066	.076
	(.025)	(.014)	(.022)	(.013)
NFOA	.090	.044	.012	.016
	(.037)	(.024)	(.034)	(.021)
FAVU	−.098	−.073	.166	.072
	(.057)	(.034)	(.058)	(.033)
Log likelihood	−220.6	−405.6	−283.4	−472.2
$\chi^2[\beta^{CR}=\beta^{MNL}; \beta^{CR}=\beta^{RV}]$	4.07	36.35	3.01	3.57

[a]Estimated standard errors are reported in parentheses.

should not be significantly different. We consider two alternative models. First, we consider a model in which each party assigns a rank of one to the arbitrator it truly prefers the most, but may rank the remaining six arbitrators strategically, that is, not in the true order of their expected utilities. The likelihood function for this model is derived from expressions for the probability that $(Y_{A(1)} > Y_{A(2)}, Y_{A(3)}, Y_{A(4)}, Y_{A(Veto1)}, Y_{A(Veto2)}, Y_{A(Veto3)})$. This is simply a multinomial logit model (*MNL*). Second, we consider a model in which each party's ranking correctly distinguishes between the four most-preferred arbitrators and the three least-preferred arbitrators, although the ranking of the top four choices may be done strategically. The likelihood function for this model is derived from expressions for the probability that $(Y_{A(1)}, Y_{A(2)}, Y_{A(3)}, Y_{A(4)} > Y_{A(Veto1)}, Y_{A(Veto2)}, Y_{A(Veto3)})$. This model shall be referred to as the rank/veto model (*RV*).[20]

In Table 4 we present estimates of the *MNL* model and the *RV* model for specification (1) in Table 3. If the rankings data do not reflect the true preferences of the parties, then the three sets of estimates will in general have different probability limits, regardless of the true underlying model. This suggests that a specification test based on the differences between the estimated parameter vectors has power against a broad range of alternative models in which the rankings do not reflect true preferences. This test should be especially powerful against strategic behavior that is closely approximated by the *MNL* or the *RV* models.[21] In constructing this test we follow Jerry Hausman (1978) who suggests a statistic of the form

$$S=(\hat{\beta}^{I}-\hat{\beta}^{II})'(\hat{V}^{I}-\hat{V}^{II})^{-1}(\hat{\beta}^{I}-\hat{\beta}^{II})$$

where the $\hat{\beta}$ are estimated parameter vectors, the $\hat{V}$ are their estimated variance-covariance matrices, and the superscripts I and II index the alternative estimators. Hausman shows

[20]It should be stressed that neither the *MNL* nor the *RV* model is presumed to be the correct structural model under strategic behavior. Both are, however, reasonable approximations to structural models that strategic behavior is likely to imply. For example, even when behaving strategically, bargainers are likely to give the top rank to one of their most preferred arbitrators. In this case, the *MNL* model would be a reasonable approximation to the underlying structural model. Alternatively, strategic behavior is unlikely to imply vetoing one's most preferred arbitrators. In this case, the *RV* model would be a reasonable approximation.

[21]In principle, the test also has power against a broad range of other specification errors relating to distributional assumption, functional form, and independence of the errors.

that, under the null hypothesis, S is asymptotically distributed as a *chi*-square random variable with degrees of freedom equal to the rank of $(\hat{V}^{I}-\hat{V}^{II})$.

Test statistics for the hypotheses $\beta^{CR}=\beta^{MNL}$ and $\beta^{CR}=\beta^{RV}$ are reported in the last row of Table 4. Since the critical value for these statistics at the 5 percent level is 11.07, none of the equality hypotheses are rejected with the single exception of $\beta^{CR}=\beta^{RV}$ for employers. Indeed, the closeness of the parameter estimates in Tables 3 and 4 is a remarkable finding that provides strong evidence that the rankings data mainly reveal the parties' true preferences. Even the two estimated parameter vectors that give rise to the χ^2 statistic of 36.35 are close in magnitude and substantively no different. The high χ^2 value appears to be the result of the two sets of estimates having virtually identical estimated variances. In other words, the statistic reflects a precise measure of a small difference; it does not reflect a large difference between the estimated β's. Overall, then, the results of this subsection provide no support for the hypothesis of strategic behavior by either party.

IV. Conclusions

This paper opens up the empirical analysis of a new area in the literature on bargaining under arbitration: the selection of arbitrators. Our major substantive findings are that 1) employers and unions distinctly prefer some arbitrators to others, 2) employer and union preferences tend to be moderately similar to each other, 3) employers tend to prefer arbitrators with training in economics whereas unions prefer arbitrators with legal training and dislike economists, 4) both parties prefer arbitrators with greater experience, 5) there is evidence that the parties' preferences are affected by arbitrators' win-loss tallies under final-offer arbitration, and 6) there is no evidence that the parties rank arbitrators strategically.

Overall, the results suggest that New Jersey's prescreening procedure for establishing its master list of arbitrators works well. Even after controlling for the arbitrators' characteristics, there is much noise in the parties' preference orderings. Nevertheless, it still seems advisable to let the parties have input into the appointment of an arbitrator, since our results indicate that they have sensible preferences that they reveal accurately. Indeed, appointment mechanisms that account for the parties' preferences tend to result in higher welfare for both parties than mechanisms that appoint arbitrators on a rotating basis.

Of course, it has been argued that arbitration should be made costly so that bargainers have an incentive to settle their disputes voluntarily (see Carl Stevens, 1966; Henry Farber and Harry Katz, 1979; Bloom, 1981). In this view, mechanisms that allow disputants to choose their arbitrator are undesirable because they reduce the (indirect) costs of arbitration. However, this negative factor must be weighed against the increased legitimacy an arbitrator will have when the parties have been involved in his (or her) appointment.

Our empirical results seem to indicate that the New Jersey mechanism for selecting arbitrators works mainly as a safety net which allows the parties to filter out arbitrators who probably should not have survived the prescreening. In other words, the extraordinary closeness of estimates derived from the complete rankings and the rank/veto model suggests that most of the information about the parties' preferences comes from the vetoed arbitrators and not from the rank order of the nonvetoed arbitrators. Indeed, it may well be true that having seven arbitrators per panel is optimal given the degree of prescreening that takes place and the nature of the parties' preferences.

Finally, the similarity of union and employer preferences for different arbitrators suggests that collective bargaining functions more cooperatively than most existing models indicate. This finding represents potentially important input into the further development of models of employer-union interactions.

REFERENCES

Ashenfelter, Orley, "Evidence on U.S. Experiences with Dispute Resolution Systems," Industrial Relations Section Working Paper No. 185, Princeton University, 1985.

______ **and Bloom, David E.**, "The Pitfalls in Judging Arbitrator Impartiality by Win-Loss Tallies Under Final-Offer Arbitration," *Labor Law Journal*, August 1983, *34*, 534–39.

______ **and** ______, "Models of Arbitrator Behavior: Theory and Evidence," *American Economic Review*, March 1984, *74*, 111–24.

Bloom, David E., "Customized 'Final-Offer': New Jersey's Arbitration Law," *Monthly Labor Review*, September 1980, *103*, 30–33.

______, "Is Arbitration *Really* Compatible with Bargaining?," *Industrial Relations*, Fall 1981, *20*, 233–44.

______ **and Cavanagh, Christopher L.**, "An Analysis of the Selection of Arbitrators," Harvard Institute of Economic Research Working Paper No. 1150, 1985.

______ **and** ______, "An Analysis of Alternative Mechanisms for Selecting Arbirators," Working Paper No. 1224, Harvard Institute of Economic Research, 1986.

Burger, Warren E., "Using Arbitration to Achieve Justice," *Arbitration Journal*, December 1985, *40*, 3–6.

Crawford, Vincent P., "On Compulsory Arbitration Schemes," *Journal of Political Economy*, February 1979, *87*, 131–60.

______, "Compulsory Arbitration, Arbitral Risk and Negotiated Settlements: A Case Study in Bargaining Under Imperfect Information," *Review of Economic Studies*, January 1982, *49*, 69–82.

Elkouri, Frank and Elkouri, Edna Asper, *How Arbitration Works*, Washington: Bureau of National Affairs, Inc., 1985.

Farber, Henry S., "An Analysis of Final-Offer Arbitration," *Journal of Conflict Resolution*, December 1980, *24*, 683–705.

______ **and Katz, Harry C.**, "Interest Arbitration, Outcomes, and the Incentive to Bargain," *Industrial and Labor Relations Review*, October 1979, *33*, 55–63.

Goldfeld, Stephen M. and Quandt, Richard E., *Nonlinear Methods in Econometrics*, New York: North-Holland, 1972.

Hausman, Jerry, "Specification Tests in Econometrics," *Econometrica*, November 1978, *46*, 1251–71.

MacDonald, Lois, "Selection and Tenure of Arbitrators in Labor Disputes," *Proceedings of First Conference on Labor*, Albany: Matthew Bender, 1948, 145–95.

McFadden, Daniel, "Econometric Models of Probabilistic Choice," in Charles Manski and Daniel McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications*, Cambridge, MIT Press, 1982.

Roth, Arthur, Kadane, Joseph B. and Degroot, Morris H., "Optimal Peremptory Challenges in Trials by Juries: A Bilateral Sequential Process," *Operations Research*, September-October 1977, *25*, 901–19.

Stevens, Carl M., "Is Compulsory Arbitration Compatible with Bargaining?," *Industrial Relations*, February 1966, *5*, 38–52.

An Investigation into the Determinants of U.S. Strike Activity

By JOSEPH S. TRACY*

This paper outlines the construction of a new panel data set of U.S. contract negotiations and strike activity. This is the first to contain data on both strikes and contract expirations. Key findings are that 15 percent of scheduled negotiations end as strikes, but strike probabilities are higher in June and lower in December; the variability, and not the level, of firm profitability affects strike activity; personal characteristics of the union workforce affect strike activity; and strikes are more likely when the local labor market is tight, but less likely when the industry labor market is tight.

To gain an understanding of why strikes occur and what factors lead to a settlement, it is important to examine data pertaining to the level at which negotiations take place, that is, the individual bargaining unit. Despite this, there has not been to date a study of the incidence and duration of strikes in the United States on a comprehensive micro data set of contract negotiations. This reflects the difficult problems involved in collecting this type of data. The purpose of this paper is to take a step in this direction by illustrating how these problems can be overcome and to present some results for the five-year period from 1973 to 1977.

One of the principal findings is the way in which a firm's profitability influences the bargaining process. The firm's level of profitability has no impact on the likelihood of a strike. However, profit volatility increases both the incidence and duration of strikes. Important scale effects were found in the data. Large firms have both lower strike probabilities and shorter strike durations. Personal characteristics of the union workforce in the industry were also important determinants of strike activity. Strike incidence is higher the more educated workers are, the younger they are, and the higher is the percentage of white workers. Finally, labor market conditions significantly affect the course of the negotiations. Strikes are less likely when industry labor markets are tight, and more likely when local labor markets are tight.

The outline of the paper is as follows. A complete description of the construction of the negotiation data is presented in Section I. Particular attention is given to discussing solutions to the problems encountered in using the Bureau of Labor Statistics (BLS) strike data. The section concludes with a discussion of the motivation for and construction of the variables to be included in the analysis. The econometric methods used to examine the data are outlined in Section II, followed by a presentation of the empirical results.

I. Construction of the Data and Variables

The bulk of the empirical work on U.S. strike activity has used aggregate time-series data. These studies typically estimate some variant of Orley Ashenfelter and George Johnson's model (1969). One of the principal difficulties with using aggregate data is controlling for the underlying number of negotiations taking place. Micro data on individual contract negotiations solve this problem in a natural way. However, a characteristic of many of the existing micro studies is that their samples include only a small number of

*Department of Economics, Yale University, New Haven, CT 06520. I thank Sherwin Rosen, Edward Lazear, and Robert Topel for their extensive comments, and also thank Gary Becker, George Neumann, John Abowd, and David Card for helpful discussions. Any errors are my responsibility.

firms (see Henry Farber, 1978; Drew Fudenberg, David Levine, and Paul Rudd, 1983; and Martin Mauro, 1982). For example, Farber's data followed ten firms while Mauro's followed fourteen. A strong point of these data sets is the long time period covered by the data. What is needed, though, is a panel data set which follows a broad spectrum of firms and unions.

A second difficulty encountered when trying to analyze U.S. strike activity is the accuracy of the strike information itself. In most studies, this information is gathered from public sources such as newspaper reports. Two potential problems exist. First, it is possible that a strike could go unreported. This may lead the researcher to miscode a negotiation as a nonstrike. Second, information on the actual number of workers involved, the duration of the strike, etc., may be subject to reporting error. This again introduces measurement error into the analysis. A concerted attempt has been made in this study to minimize both of these problems. The techniques used will be outlined below.

The focus of this study will be on strikes that occur during renegotiations of contract terms. This excludes, for example, organizational strikes and sympathy strikes from the analysis. The omitted categories comprise about 40 percent of all major strikes.[1] This bargaining process can be initiated in one of two ways. The first is a scheduled negotiation. This can occur either at the expiration of the current contract, or at an agreed-upon reopening of the contract. The second manner is an unscheduled negotiation due to an unanticipated reopening of the contract. This latter type of negotiation occurs infrequently and usually in response to a dramatic development that requires immediate attention.

This study deals exclusively with scheduled negotiations. This decision was made because of the difficulty in obtaining reliable information on unscheduled reopenings. Information on scheduled negotiations is available from the BLS. This information is based on a file of union contracts which the BLS maintained. For recent years, they had fairly extensive coverage of major contract, that is, contracts covering 1,000 or more workers. The expiration and reopening dates for all of these contracts were published annually in the bulletin *Wage Calendar*. Information on smaller bargaining units is available only in unpublished form and is not as comprehensive in coverage. Consequently, only major contracts are included in the sample. While major contracts account for less than 50 percent of all contracts, they cover roughly 90 percent of all union workers. The sample was further restricted to manufacturing contracts.

For each contract, the BLS lists the year and month it expires, the name of the firm(s) and the union(s) comprising the bargaining pair, the number of workers covered by the contract, the 2-digit SIC classification for the major product line affected, and the state or region involved. In addition, a contract identification number is assigned which allows you to follow that bargaining pair through each successive contract negotiation. The day of the expiration, the contract length, and the 4-digit SIC classification were found in unpublished listings provided by the BLS. The contract identification numbers made it possible to merge this additional information in with the published expiration data.

The BLS also collected extensive data on U.S. strike activity. This information was summarized annually in their bulletin *Work Stoppages*. The collection process began with an unpublished weekly summary of strikes in progress, *Industrial Relations Facts* (*IRF*), which the BLS gathered from public sources. Each company listed as being involved in a strike was contacted and a request was made for verification of the information given in the public source. To increase their response rate and to insure the accuracy of the information provided, the BLS pledged confidentiality over the use of this strike data. As a result, the BLS has only released this data with the names of the firm and the union removed from each record.

As a consequence of this confidentiality issue, the only published data from the BLS strike file that identifies the names of the

[1] These estimates are based on work stoppage data provided by George Neumann.

parties to the negotiations are for contracts with 10,000 or more workers. This is the fundamental roadblock confronting anyone attempting to construct a data set at the bargaining unit level of observation. Restricting the analysis to contracts involving 10,000 or more workers would significantly reduce the sample size and coverage. Including all contracts of 1,000 or more workers would seem to preclude using the most accurate information available. Relying on the *IRF* data would probably be adequate for studies of strike incidence among major contract negotiations, since it is unlikely that many major strikes would go unreported. However, not having access to the BLS work stoppage data could be a more serious problem for studies of strike durations.

The best solution to this problem seems to be to try and circumvent the difficulties raised by the confidentiality issue. If the names of the firm(s) and the union(s) could be recovered and reinserted onto the BLS strike data, then the sample could include all major contracts and still use the cleaned-up BLS strike information. Two sources of information are available for this identification effort. The primary source is the strike listings of the *IRF*. A secondary source was the BLS publication *Current Wage Developments* (*CWD*). The purpose of *CWD* is to report on the major changes in the contract provisions following a negotiation. However, *CWD* indicates for some contracts that the settlement was preceded by a strike of some specified duration. Presumably, this information is based again on secondary sources and not the actual BLS strike data.

I received from the BLS a set of the *IRF* covering the period from 1973 to current. For each strike listed, I followed the strike through each weekly issue from its start up to the settlement. This provided a single observation for each strike which contained all of the information that is also reported on the BLS strike tape, with the exception of a contract status and major issue variable. The strike listing generated from the issues of the *IRF* was merged with a similar listing complied from issues of the *CWD*. I then matched the strikes from the BLS work stoppage tape to the strikes from this combined public listing using the overlap in information. Care had to be taken since the information in the public listing was subject to reporting error. Using this procedure, I was able to recover the names of the firms involved in over two-thirds of the strikes during the period from 1973 to 1977.

Having reinserted as many names as possible, I then selected a subsample of strikes relevant to my analysis. These selections were necessary to make the strike sample conform with the negotiation sample. Strikes were kept if they took place at an expiration or reopening of a contract, and if they involved 1,000 or more workers. Separating out the strikes by type is possible since the work stoppage data includes a contract status variable that indicates whether contract terms were under negotiation at the time of the strike. A strike satisfying the above criteria was kept even if no match had been found in the public strike listing. The reason for this is that, even without the names of the firm and the union, it is sometimes possible to match a strike with its corresponding expiration using common information on detailed SIC classification, region, number of workers, and dates.

For the period from 1973 to 1977, the total sample contains 2,100 contract negotiations for which detailed strike information is available. The sample consists of 1,130 distinct bargaining pairs, 392 firms, and 75 unions. A total of 120 3-digit industry classifications and 45 states are represented in the data.[2] Tables 1 and 2 show the distribution of these contracts and strikes by year and by month. They illustrate the uneven distribution of negotiations across both years and months within a year. This underscores the point made earlier concerning the importance of being able to control for the amount of negotiating activity.

[2] Several contract negotiations had to be dropped from the estimation due to missing information on variables used in the analysis. The sample used in the estimation contains 1,319 contract expirations and reopenings involving 358 firms and 61 unions. Currently, the sample is being expanded to include all major manufacturing and nonmanufacturing negotiations from 1970 to the present.

TABLE 1—DISTRIBUTION OF CONTRACT EXPIRATIONS AND STRIKES BY YEAR

Year	Number of Contract Expirations	Number of Strikes	Strike Frequency
1973	214	40	18.69
1974	327	65	19.88
1975	187	19	10.16
1976	248	35	14.11
1977	343	39	11.37
Total	1,319	198	15.01

TABLE 2—DISTRIBUTION OF CONTRACT EXPIRATIONS AND STRIKES BY MONTH

Month	Number of Contract Expirations	Number of Strikes	Strike Frequency
January	99	14	14.14
February	65	11	16.92
March	102	10	9.80
April	122	22	18.03
May	148	20	13.51
June	150	34	22.67
July	107	12	11.21
August	177	25	14.12
September	129	20	15.50
October	117	18	15.38
November	49	10	20.41
December	54	2	3.70
Total	1,319	198	15.01

In the remainder of this section I explain why I selected the variables included in the analysis. I will also provide details concerning the construction of these variables. The first set of variables takes advantage of the fact that we know the firm involved in the bargaining. Two sources of firm-specific data were used. The CRSP data provides security price information and the COMPUSTAT data provides accounting information. Both data bases use the same firm identification number called a CUSIP number.

The only difficulty encountered in adding these CUSIP numbers to the data set was handling mergers and takeovers. When these occur, firm-level information may no longer be available for one or possibly both firms involved. However, the BLS will typically continue to list the old firm names in the expiration data. In these cases, the appropriate new firm name was obtained from a directory of firms and the CUSIP number for that firm was added to the data.

The first aspect the firm attempts to control for is its recent profit performance. Several micro studies have included measures of the firm's profitability. However, the level of profitability may not be the only relevant feature. Harold Grubert (1968) argued that instability in firm profits might lead to increased strike activity for two reasons. The first is that this volatility might make the union leadership less certain about the firm's willingness to make concessions during the bargaining. Secondly, Grubert argued that to the extent that management has better information about the firm's future profitability, "... it may be necessary for the union to threaten a strike or to begin one in order to force the company to reveal the level of profits it really expects" (p. 23). Grubert did not explicitly model how a strike might allow the union to infer the firm's private information. However, recent game-theoretic models of bargaining have been based exactly on such an idea (see Peter Cramton, 1982; Fudenberg and Jean Triole, 1981; Beth Hayes, 1984; Joel Sobel and Ichiro Takahashi, 1983; and my 1984 dissertation).

Grubert measured profit instability as the sum of the absolute deviations of annual profits from trend over the past five years scaled by the firm's employment. He found that this measure of volatility had a positive effect both on the profitability and the conditional duration of a strike with the later effect significant at the 0.025 level (p. 40). For the level of the firm's performance, I chose to use the rate of return on the firm's stock for the year preceding the contract expiration. The volatility measure used is the standard deviation of the firm's daily stock return. This is calculated on a year of daily trading data ending six months prior to the contract expiration.[3]

[3]The reason for not including trading data right up to the contract expiration is that speculation over the upcoming contract negotiations will begin to occur as this expiration date approaches. This speculation will induce variability into the stock returns that need not reflect any uncertainty about the firm's future demand conditions. Instead, this variability may simply reflect uncertainty about the division of future profits between the union and the stockholders (see John Abowd, 1985).

The next two firm-specific variables attempt to control for the firm's ability to self-insure against the event of a strike. C. Lawrence Christenson (1953), in discussing the effects of the coal strikes, mentions two basic ways for firms to "offset" the interruption in the flow of union labor services. The first method is to use intertemporal substitution in production; that is, the firm builds "buffer" inventories prior to the start of the negotiations. The second method is for the firm to attempt to continue production at a reduced rate during the course of the strike.

Estimating a firm's buffer stock of inventory prior to the expiration of its contract is a difficult task. The COMPUSTAT data provides inventory data at the firm level. While COMPUSTAT asks for inventory by stage-of-process, most firms only report total inventory levels. Consequently, in order to prevent the sample size from being significantly reduced, raw materials, work-in-progress, and final goods must be assumed to provide identical contributions toward the firm's insurance efforts.

The other limitations of the inventory data are that only year-end figures are given and all product lines are aggregated together. Ideally, we would like to observe the firm's inventory position close to the expiration date and we would like to focus just on the product lines that may be potentially affected by a strike. The proxy used for the firm's buffer stock is the percentage change in its inventory-to-sales ratio for the year preceding the contract expiration. The inventory figures were scaled by the firm's sales in order to account for normal inventory growth due to sales growth.

The second method available to the firm to offset the costs of a strike is to attempt to maintain production during the strike. The ability of the firm to continue production may in part be determined by how capital intensive the production technology is in that industry. Firms in highly capital-intensive industries may be able to train their managers to continue operations at a reduced rate. An example of this was the nationwide telephone strike. Due to the high degree of automation in the telephone industry, many types of services continued throughout the strike. The firm's capital-labor ratio was calculated using the previous year's net plant and equipment and total firm employment.

The final firm-specific variable included in the analysis is a measure of firm size. Significant scale effects have been found in studies of wage determination. This indicates that the structure of internal labor markets within a firm may differ in important ways with the size of the firm. These differences may also affect the bargaining process. Firm size could be measured with either sales, capital stock, or employment. Since the latter two were used to form the capital-labor ratio, sales were chosen as the size measure.

The next set of variables attempts to control for the personal characteristics of the union workforce. Ideally, we would like to have information on the union workers actually in the bargaining unit. Since this information is not available, I constructed measures for the union workers in the same industry as the bargaining unit. Individuals who were working full time and who were covered by a union contract were selected from May *Current Population Survey* tapes from 1973 to 1977. These individuals were pooled and then sorted by 2-digit industry classifications. Industry averages for the age, education, percent male, and percent white were then calculated.

Two variables were added to the data to control for differences in industry structure. First, it is possible that the presence or absence of monopoly rents in an industry may significantly affect the bargaining process. A measure of potential monopoly rents that has been used extensively in the past (as well as debated over) is the concentration ratio. Specifically, this ratio is the percent of the total sales in a 4-digit industry classification that is accounted for by the four largest firms. The second variable is the percent of the industry employment that is unionized. The motivation for this variable is that higher unionization rates may place the union in a stronger relative bargaining position. Econometric studies of union wage effects often find a positive and significant effect for this variable (see H. Gregg Lewis, 1983). Estimates of unionization rates at a 3-digit industry level were taken from the work of Richard Freeman and James Medoff (1979).

Labor market conditions at the time of the negotiations may also be important determinants of strike activity. Ashenfelter and Johnson argue that "... during periods of low unemployment there will be decreased opposition among the rank and file to a militant course of action since there will be part-time job opportunities available for potential workers" (p. 40). The national unemployment rate has been the measure used in many previous studies. Almost unanimously, the finding among studies is that this unemployment rate has a negative and significant effect on the amount of strike activity.

The potential exists in this data set to more fully characterize the labor market conditions. For each negotiation, we know both the industry and the region affected. It would be interesting, then, to separately control for the industry and the regional labor market conditions. The only state unemployment rates going back to the early 1970's are constructed from state Unemployment Insurance claims. While attempts have been made to remove any inconsistencies due to the differences in state Unemployment Insurance laws, I decided not to use this data. Instead, both the industry and the local labor market conditions will be measured in terms of residuals from trend employment.

The industry trend regressions were estimated using quarterly 3-digit employment data for the period 1970–81. The local regressions were based on state or regional employment for the same time period.[4] The specification estimated was

$$(1)\quad \ln E_{it} = \beta_{i0} + \beta_{i1}t + \sum_{j=1}^{3} \delta_{ij} Q_j + U_{it}$$

$$U_{it} = \Phi(L) U_{it-1} + \epsilon_{it}$$

where $\ln E_{it}$ = log quarterly employment in industry or region i at time t; Q_j = dummy variable for the jth quarter; $\Phi(L)$ = distributed lag polynomial; and ϵ_{it} = white noise.

The order of Φ was chosen so that the ϵ process showed no serious indications of departure from white noise. In most cases, a first- or second-order polynomial was sufficient. A potential feedback problem exists if the actual residuals are used in the analysis. The BLS gathers the monthly employment figures by counting the number of workers on payrolls as of the second week of the month. Workers on strike at this time are not added into the figures. This introduces a negative correlation between the actual level of strike activity and the estimated residual.

The autoregressive structure of the estimated residuals provides a method for avoiding this feedback problem. The current residual can be decomposed into a predicted and an unpredicted component. The predicted component is calculated using the $\Phi(L)$ polynomial and past employment residuals. Consequently, the predicted component of the current residual should be free of any significant correlation with the actual extent of strike activity in that quarter.

The employment growth rates from the trends regressions will be used to control the long-term trends in the industry and locality. Several other variables that need no explanation will also be tested. The next section explains the econometric methods used in the analysis and presents the empirical results.

II. Empirical Specification and Results

Two alternative estimation strategies exist for testing a variable's impact on the probability and duration of a strike. The first approach is to jointly estimate these effects using a Tobit model. The alternative is to estimate separate models for the probability and the conditional duration. The Tobit model builds in the assumption that if a variable increases the likelihood of a strike it also increases the conditional duration. While this may be a reasonable restriction, the second estimation strategy allows the data to indicate this rather than assuming it be be true. For this reason, the second approach will be used here.

[4] When a contract involves two or more states from different regions, the BLS assigns an interstate code. If these individual states could be identified, then the residual used is a population weighted average of the state residuals.

I assume that the probability of a strike occurring during the contract negotiations between union i and firm j at time t is given by the logistic function

$$(2)\qquad Pr_{ijt} = 1/\left(1+\text{Exp}(-X_{ijt}\beta^s)\right).$$

The implied marginal effect of a variable on the probability of a strike is given by the function

$$(3)\qquad \frac{\partial Pr}{\partial X_j} = \beta_j^s \frac{\text{Exp}(-X\beta^s)}{[1+\text{Exp}(-X\beta^s)]^2}.$$

The transition from a strike to a settlement is modeled by a hazard function, $\lambda(t; X)$. The choice of the hazard function uniquely determines the probability distribution function for the conditional strike durations. The probability of observing a strike of duration t^* days conditional on its occurrence is

$$(4)\qquad f(t^*; X) = \lambda(t^*; X)\text{Exp}\left[-\int_0^t \lambda(t; X)\, dt\right].$$

The manner in which time and the exogenous variables affect the hazard rate must be specified. A hazard function exhibits "duration dependence" if $\partial\lambda(t; X)/\partial t \neq 0$. In particular, when $\partial\lambda(t; X)/\partial t > 0$, positive duration dependence exists. In this case, the longer a strike continues, the more likely it is that the strike will be settled in the next interval of time. If no duration dependence exists, then the conditional strike durations follow an exponential distribution.

I use a form for the hazard function that allows for any monotonic duration effect:

$$(5)\qquad \lambda(t; X) = \lambda\gamma(\lambda\gamma)^{\gamma-1}h(X),$$

so that

$$(6)\qquad \partial\lambda(t; X)/\partial t = \lambda\gamma(\gamma-1)(\lambda t)^{\gamma-2}h(X) \gtreqless 0 \text{ as } \gtreqless 1.$$

This allows the data to select the type of duration effect through the estimated value of *gamma*. The parameter γ is called the "baseline" hazard.

The remaining choice is the form for the function $h(X)$. A widely used functional form is the exponential: $h(X) = \text{Exp}(X\beta^d)$. This choice for $h(X)$ gives the "proportional" hazard model. Let X_t denote the value of the exogenous variables at the outset of the negotiations. Assuming that these variables are held constant throughout the strike, then the probability that a strike starting at time t continues for t^* days is

$$(7)\qquad f(t^*; X_t) = \lambda(t^*; X_t)\text{Exp}\left[\left(-\lambda(t^*; X_t)t^*\right)/\gamma\right].$$

The assumption that the exogenous variables remain constant throughout a strike, though, is unreasonable given that over half of the strikes in the sample continue beyond the quarter in which the contract expired. The longest strike lasted for a total of seven quarters. Consequently, the hazard function will incorporate variations in the industry and the local employment residuals as a strike enters a new quarter.

Partition the vector of exogenous variables, X_t, into a subvector that remains constant during a strike, X_{1t}, and a subvector that can vary from quarter to quarter, X_{2t}. Let t_k denote the number of days from the outset of the strike to the end of the kth quarter if the strike continues beyond that quarter; otherwise, t_k is the total duration of the strike. The probability of a strike starting at time t, lasting t^* days, and involving k quarters is

$$(8)\qquad f(t^*; X_t) = \lambda\gamma(\lambda t^*)^{\gamma-1}\text{Exp}\left[X_{1t}\beta_1^d\right]\text{Exp}\left[X_{2t_k}\beta_2^d\right] \times \text{Exp}\left[-\lambda^\gamma \text{Exp}\left(X_{1t_1}\beta_1^d\right)\left\{t_1^\gamma \text{Exp}\left(X_{2t_1}\beta_2^d\right) + \sum_{j=2}^{k}\left(t_j^\gamma - t_{j-1}^\gamma\right)\text{Exp}\left(X_{2t_j}\beta_2^d\right)\right\}\right].$$

The implied marginal effect of a variable on

the conditional strike duration is

$$(9)\ \partial E(D|S)/\partial X_j = \frac{-\beta_j^d}{\lambda\gamma}\frac{\Gamma(1+1/\gamma)}{\left[\mathrm{Exp}(X\beta^d)\right]^{1/\gamma}}$$

Prior to estimation, the data were standardized by subtracting out the variable means and dividing by their standard deviations. These means and standard deviations are presented in Table 3. Tables 4 and 5 give the estimated coefficients from the logistic and hazard models as well as the implied marginal effects. The hazard marginal effects are measured in calendar days.[5]

Specification (2) differs from specification (1) in each table in that it includes fixed effects for eight major unions. Several variables have been constructed to capture differences among firms involved in the bargaining. However, no similar variables capture heterogeneity among unions. An example of such a union-specific variable that has been used in previous studies is a measure of potential strike benefits. Farber used the union's national strike fund balance per member while Grubert used the monthly contribution per member to the national fund. Farber found no significant effect for his proxy, while Grubert reported that his proxy had a positive and significant effect of the probability of a strike and a negative and significant effect on the duration of a strike.

Given the difficulty in obtaining estimates of these strike benefits and the inconclusive findings to date, the approach taken here is to not attempt to specify the source of the union heterogeneity. To the extent that these differences remain roughly constant both across bargaining pairs and through time, then their influence can be captured by a simple fixed effect. The choice of which unions to include a fixed effect for was dictated by the need for a sufficient number of contract negotiations and strikes involving that particular union. Consequently, a fixed effect was estimated for a union if at least five strikes involved that union. Eight unions comprising 51 percent of the negotiation sample and 74 percent of the strike sample satisfied this selection rule. Table A1 gives these fixed-effect estimates. (A table showing the distribution of each union's negotiations across major industry classifications can be obtained from the author.)

Turn now to the results given in Tables 4 and 5. Consider first the impact of the firm-specific variables. The firm's performance as measured by the rate of return on its stock has no effect of the likelihood of a strike. Conditional on a strike occurring, a one-standard-deviation increase in the firm's rate of return results in slightly over a five-day reduction in the expected duration. However, this effect is not very precisely measured and

TABLE 3—UNCONDITIONAL SAMPLE MEANS AND STANDARD DEVIATIONS

Variable	Mean	Standard Deviation
Rate of Return on Stock	7.2583	38.0756
Volatility of Stock Returns	0.0204	0.0077
Net Sales	3,598.2704	6,665.6834
Change in Inventory/Sales	−2.2368	16.1865
Capital-Labor	23.0301	30.8244
Average Age	39.6963	1.1031
Average Education	11.9631	0.4174
Percent White	87.0558	3.7908
Percent Male	80.3392	14.1186
Concentration Ratio	45.9378	21.0837
Union Coverage Rate	42.6122	12.4427
Industry Predicted Employment Residual	0.0816	5.0055
Local Predicted Employment Residual	−0.6690	4.0798
Industry Employment Growth Rate	0.1284	0.4494
Local Employment Growth Rate	2.1749	1.1450
Conditional Duration	50.0000	64.9289

[5]Table 4 also reports "pseudo" R^2 statistics for each specification. This R^2 is calculated as follows:

$$R^2 \equiv \left(1-(L_\beta/L_\Omega)^{2/N}\right)/\left(1-(L_\Omega)^{2/N}\right),$$

where L_β = maximized value of the unrestricted likelihood function, L_Ω = maximized value of the likelihood function restricted to an intercept term, and N = sample size. This measure was proposed by John Cragg and Russell Uhler (1970).

TABLE 4—LOGISTIC MODEL

Variable	Logistic Coefficient (1)	Marginal Effect (1)	Logistic Coefficient (2)	Marginal Effect (2)
Intercept	−2.02396		−2.45879	
	(−20.96)		(−15.92)	
Rate of Return on Stock	0.02819	0.00291	0.00897	0.00088
	(0.34)	(0.34)	(0.10)	(0.10)
Volatility of Stock Returns	0.24225	0.02497	0.20536	0.02010
	(2.71)	(2.70)	(2.22)	(2.22)
Net Sales	−0.40221	−0.04146	−0.32898	−0.03220
	(−3.03)	(−3.08)	(−2.45)	(−2.47)
Change in Inventory/Sales	−0.06434	−0.00663	−0.02306	−0.00226
	(−0.77)	(−0.77)	(−0.26)	(−0.26)
Capital-Labor	0.23032	0.02374	0.19737	0.01932
	(1.72)	(1.72)	(1.40)	(1.41)
Average Age	−0.62140	−0.06313	−0.41975	−0.04108
	(−5.18)	(−5.41)	(−3.03)	(−3.11)
Average Education	0.37123	0.03827	0.26119	0.02556
	(2.71)	(2.77)	(1.72)	(1.75)
Percent Male	−0.09032	−0.00931	−0.08125	−0.00795
	(−0.78)	(−0.78)	(−0.58)	(−0.58)
Percent White	0.36811	0.03795	0.44929	0.04398
	(3.04)	(3.09)	(3.17)	(3.27)
Concentration Ratio	0.26354	0.02717	0.22710	0.02223
	(2.65)	(2.66)	(2.14)	(2.14)
Union Coverage Rate	0.20853	0.02150	0.21608	0.02115
	(1.60)	(1.62)	(1.38)	(1.40)
Industry Predicted Employment Residual	−0.20480	−0.02111	−0.22903	−0.02242
	(−2.31)	(−2.32)	(−2.43)	(−2.44)
Local Predicted Employment Residual	0.47323	0.04878	0.51824	0.05072
	(4.57)	(4.65)	(4.69)	(4.80)
Industry Employment Growth Rate	0.10218	0.01053	0.14674	0.01436
	(1.07)	(1.07)	(1.42)	(1.42)
Local Employment Growth Rate	0.17486	0.01802	0.20055	0.01963
	(2.02)	(2.02)	(2.23)	(2.24)
Log Likelihood	−499.737		−476.316	
Pseudo R^2	0.15		0.20	
$N = 1{,}319$				

Note: Specification (1) contains no union fixed effects and specification (2) contains fixed effects for eight unions.

disappears when the union fixed effects are introduced. On the other hand, greater volatility of the firm's stock returns increases both the probability and conditional duration of a strike. A one-standard-deviation increase in the measure of volatility results in over a $2\frac{1}{2}$ percent increase in the strike probability and nearly a seven-day increase in the expected duration. Controlling for the union fixed effects reduces both marginal effects, but the incidence effect remains significant.

The size of the firm as measured by its previous year's sales has a large and significant effect on both the likelihood and duration of a strike. A one-standard-deviation increase in sales reduces the probability of a strike by 4 percent and the expected duration by over two weeks. These effects remain significant even when the union fixed effects are included. The measure for the firm's buffer inventory has no effect on either measure of strike activity. Finally, there is some indica-

TABLE 5—PROPORTIONAL HAZARD MODEL

Variable	(1) Hazard Coefficient	(1) Conditional Marginal Effect	(2) Hazard Coefficient	(2) Conditional Marginal Effect
Rate of Return on Stock	0.11020	−5.17477	0.05862	−2.54764
	(1.55)	(−1.52)	(0.78)	(−0.78)
Volatility of Stock Returns	−0.14723	6.91371	−0.14620	6.35402
	(−1.75)	(1.77)	(−1.61)	(1.65)
Net Sales	0.31198	−14.65060	0.36441	−15.83820
	(2.41)	(−2.51)	(2.76)	(−2.76)
Change in Inventory/Sales	0.03048	−1.43129	0.05473	−2.37867
	(0.37)	(−0.37)	(0.63)	(−0.62)
Capital-Labor	−0.21169	9.94070	−0.22475	9.76797
	(−1.63)	(1.66)	(−1.64)	(1.68)
Average Age	0.03843	1.80470	0.09489	−4.12425
	(0.33)	(−0.34)	(0.66)	(−0.67)
Average Education	0.05622	−2.63997	0.17423	−7.57257
	(0.43)	(−0.43)	(1.16)	(−1.14)
Percent Male	0.04595	−2.15769	0.02800	−1.21711
	(0.40)	(−0.40)	(0.20)	(−0.20)
Percent White	0.26586	−12.48470	0.15704	−6.82541
	(1.93)	(−1.89)	(1.04)	(−1.04)
Concentration Ratio	0.07479	−3.51197	0.00045	−0.01975
	(0.77)	(−0.76)	(0.00)	(−0.00)
Union Coverage Rate	0.22786	−10.70020	0.17704	−7.69461
	(1.45)	(−1.42)	(0.92)	(−0.92)
Industry Predicted Employment Residual	0.01724	−0.80967	−0.01225	0.53267
	(0.19)	(−0.19)	(−0.14)	(0.14)
Local Predicted Employment Residual	0.27544	−12.93450	0.30545	−13.27560
	(2.72)	(−2.54)	(2.86)	(−2.60)
Industry Employment Growth Rate	0.12794	−6.00782	0.15694	−6.82111
	(1.23)	(−1.21)	(1.46)	(−1.44)
Local Employment Growth Rate	0.07071	−3.32044	0.17670	−7.67962
	(0.84)	(−0.83)	(1.96)	(−1.88)
Lambda	0.01923		0.02039	
	(10.78)		(6.87)	
Gamma	1.07641		1.13320	
	(18.14)		(17.88)	
Log Likelihood	−952.652		−944.129	
$N = 198$				

Note: Specification (1) contains no union fixed effects and specification (2) contains fixed effects for eight unions.

tion that highly capital-intensive industries have higher strike probabilities and longer expected durations.

The personal characteristics of the union workforce seem to be important determinants of strike activity. Increasing the average age of the union workers by 1.1 years is associated with a dramatic 6 percent decline in the strike probability. While controlling for the union fixed effects lowers this estimate to 4 percent, it is still highly significant. In addition, increasing the average education level by 0.42 years is associated with nearly a 4 percent increase in the incidence of strikes. While both age and education play important parts in the logistic function, neither seems to affect the expected conditional duration of a strike. The percent of the union workforce that is male had no significant effect on the amount of strike activity. On the other hand, increasing the percent that is white leads to more frequent but shorter strikes. Including the union fixed effects increases the marginal effect of the racial

composition variable on the strike probability, but cuts the duration marginal effect and its significance level by almost half.

Industry structure as measured in this study does not seem to be a major factor in determining strike activity. The degree of concentration in the industry does have a positive and significant effect on strike probabilities, yet it has no effect on expected durations. Increases in the union coverage rate tends to result in more frequent but shorter strikes. However, neither of these effects is measured with great precision.

An interesting result comes out of looking at both the industry and local labor market conditions. Above-average predicted employment in the industry significantly reduces the likelihood of a strike, while similar conditions in the locality significantly increase this likelihood. The marginal effect for the local labor market conditions is also twice the magnitude of the industry effect. Both marginal effects are increased in size and significance when we control for the union fixed effects. While each variable plays an important role in determining strike incidence, only the local conditions also affect the expected duration of a strike. A one-standard-deviation increase in the local predicted employment residual is associated with a thirteen-day reduction in the expected conditional duration. In addition, higher employment growth rates in the locality are also associated with higher strike frequencies.

The estimate for *gamma* in specification (1) does not significantly differ from one. This would tend to indicate that no duration dependence exists in the data. However, if unobserved heterogeneity among the bargaining pairs is present, then it is easy to show that this estimate for *gamma* will be biased downwards. One possible source of this heterogeneity could be differences among unions. Notice that including the union fixed effects does increase the estimate of *gamma* significantly above one. This implies that the conditional settlement probability does increase with the length of the strike.[6]

Several other variables were also tested that are not reported in Tables 4 and 5. The sample contains both contract reopenings as well as renegotiations. It is possible that the likelihood of a strike differs significantly between these two types of bargaining situations. A dummy variable for a contract reopening was added to the basic specification. The logistic coefficient was -1.03416 with a standard error of 0.74682. The point estimate indicates a lower strike probability for reopeners of around 10 percent, but this effect is not measured with much precision. In addition, it may be possible that bargaining is affected by how long it has been since the last contract negotiation. The logistic coefficient for contract length measured in months was 0.02119 with a standard error of 0.10201. There is no evidence, then, that contracts with longer durations are more or less difficult to renegotiate.

Grubert hypothesized that strikes would be less likely when either a small fraction or a large fraction of the firm's employment was involved. His basic argument is as follows: "If the labor share is extremely high or extremely low, strikes are unlikely because the party with the larger share will be willing to make acceptable concessions to the other party. That is, there will be a big loser who will readily give in to the other side" (p. 16). It is possible to do a simple test for this inverted U-shaped effect since both the number of workers covered by this contract and total firm employment are available. The finding was that increasing the employment share has a negative and diminishing impact on the probability of a strike throughout the range of shares in the data. While no evidence for Grubert's hypothesis was found, this measure does not account for factors such as possible spillover effects.

Recall that Table 2 listed the sample strike frequencies by month. These ranged from a high of 22.67 in June to a low of 3.70 in

[6]John Kennan (1986), and Alan Harrison and Mark Stewart (1985), estimate strike duration models which allow for more general duration dependence effects. Kennan uses U.S. data and finds that the hazard first decreases and then increases. Harrison and Steward use Canadian data and find that the hazard slowly increases throughout the first 99 days.

TABLE 6—TEST FOR SEASONALITY OF STRIKES

Month	Difference in Strike Frequency	Logistic Coefficient	Difference in Strike Probability
January	−1.24	0.42371	4.58
		(1.00)	(0.97)
February	1.54	0.40068	4.30
		(0.89)	(0.84)
March	−5.88	−0.20232	−1.71
		(−0.46)	(−0.46)
April	2.65	0.28564	2.93
		(0.76)	(0.76)
May	−1.87	0.08912	0.84
		(0.24)	(0.24)
June	7.29	0.61550	7.16
		(1.78)	(1.83)
July	−4.17	−0.42174	−3.26
		(−1.00)	(−1.00)
August	−1.26	0.37929	4.03
		(1.02)	(1.02)
September	0.12	−0.11271	−0.99
		(−0.30)	(−0.30)
October	–	–	–
November	5.03	0.14236	1.38
		(0.31)	(0.30)
December	−11.68	−1.41208	−7.50
		(−1.81)	(−2.36)

December. This variability does not in itself indicate that there is seasonality to strike probabilities. There may be monthly variation in other factors in the model that would account for this result. To test for this seasonality, monthly dummy variables were added to the logistic function. October was selected to be the omitted month. Table 6 gives the differences in monthly strike frequencies from October, the logistic coefficient, and the implied differences in monthly strike probabilities from October. Only June and December have strike probabilities that differ significantly from October once other factors are controlled for.

As an aid in seeing how well the variables tested in this study explain interindustry variation in strike activity, Table 7 gives the industry strike frequencies and average strike durations. Strike probabilities and expected durations were then calculated for each negotiation in the sample. The model without the union fixed effects was used in these calculations. These were then averaged by industry and included as well in Table 7. For industries with a large number of observations, the two sets of figures are in fairly close agreement.

III. Summary of the Findings

Firm-specific factors are key determinants of strike activity. It is not the rate of return on a firm's stock, but rather its volatility that affects the bargaining process. This suggests that the asymmetric information theories of strikes should be carefully examined. A connection may exist between this measured variability and the benefit to the union from trying to infer from the firm information about future demand conditions. Large firms were found to have significantly less strike activity. This result should also be examined in future work on bargaining models. Finally, capital-intensive technologies tend to increase the frequency and duration of strikes. If higher capital-labor ratios indicate a greater ability for the firm to maintain production during a strike, then this finding is consistent with the view that strikes will be used more when their joint costs are smaller (see Melvin Reder and George Neumann, 1980).

While the personal characteristics of the union workforce have not been included in previous strike studies, they do have important effects on the bargaining process. Older and less-educated workers tend to be less involved in strike activity. Greater proportions of nonwhites in the workforce also reduces the use of strikes. Clearly, these factors should be incorporated into models of bargaining.

Studies of union and nonunion wages have found that the union wage differential increases with the rate of unionization in that industry. These wage gains, though, are not accompanied by any significant difference in the number of strikes. More concentrated industries do experience a higher incidence of strikes but no difference in expected durations. These findings suggest that industry structure is not a primary determinant of strike activity.

The contrasting effects of the industry and the local labor market conditions may also be consistent with the joint cost view of strikes. Above-average industry conditions

TABLE 7—AVERAGE AND EXPECTED VALUES FOR MEASURES OF STRIKE ACTIVITY BY INDUSTRY

Industry	N	Strike Frequency	Expected Strike Probability	Average Conditional Duration	Expected Conditional Duration
Food	94	0.05319	0.05952	94.40	60.22
Tobacco	14	0.07143	0.02291	42.00	218.06
Textile	23	0.00000	0.02269	–	–
Apparel	13	0.00000	0.01180	–	–
Lumber	20	0.10000	0.09654	98.00	86.41
Furniture	14	0.07143	0.09983	47.00	94.20
Paper	112	0.07143	0.14094	29.00	45.84
Printing	10	0.10000	0.13640	101.00	57.81
Chemicals	121	0.06612	0.10754	79.38	69.71
Petroleum	47	0.08511	0.03339	47.25	53.24
Rubber	41	0.46341	0.28316	52.84	61.99
Leather	18	0.11111	0.10289	85.00	74.11
Stone, Clay and Glass	55	0.05454	0.08504	37.33	36.58
Prim. Metal	157	0.08917	0.06812	37.50	39.60
Fab. Metal	50	0.20000	0.16051	53.20	44.81
Mach. except Elec.	152	0.28289	0.27766	39.65	39.00
Elec. Eq.	161	0.19876	0.21541	55.38	48.35
Trans. Eq.	182	0.20330	0.20176	52.62	54.91
Prof. Instr.	25	0.28000	0.20688	28.43	51.22
Misc.	10	0.10000	0.07585	20.00	97.03

may indicate that it is expensive for production to be halted due to a strike. On the other hand, above-average local conditions may indicate that union workers have good opportunities outside the firm during a strike. In any event, the findings point out the importance of controlling separately for each type of labor market effect.

Finally, the significance of some of the union fixed effects suggest that it is important to attempt to characterize specific ways in which unions may differ. The distribution of union contracts by industry illustrate that some of the union effects must be interpreted with caution. For example, both the United Rubber Workers and the Oil, Chemical, and Atomic Workers unions are both heavily concentrated in a single industry and comprise a large fraction of contracts in that industry. Consequently, it is impossible to say whether this is a union or an industry fixed effect. On the other hand, the Marine and Shipbuilding Workers, while being entirely concentrated in the transportation equipment industry, comprise only 6 percent of the total contracts in that industry. The union fixed effect in this case is clearly not simply an industry effect in disguise.

TABLE A1—UNION FIXED EFFECTS

Union	Logistic Coefficient	Hazard Coefficient
Electrical Workers (IBEW)	0.12194	−0.73828
	(0.27)	(−1.75)
Machinists	0.45918	−0.49669
	(1.45)	(−1.68)
Marine and Shipbuilding Workers	2.67857	−0.49938
	(3.98)	(−1.05)
Rubber Workers	1.83145	0.30769
	(4.19)	(0.78)
Steelworkers	0.29866	0.06455
	(1.04)	(0.23)
Electrical Workers (IEU)	0.91477	0.40752
	(2.08)	(0.99)
Oil, Chemical, and Atomic Workers	1.22526	−0.73493
	(2.52)	(−1.48)
Auto Workers	1.01437	0.23874
	(3.70)	(0.94)

REFERENCES

Abowd, John M., "Collective Bargaining and the Division of the Value of the Enterprise," Working Paper, University of Chicago, October 1985.

Ashenfelter, Orley and Johnson, George E., "Bargaining Theory, Trade Unions, and Industrial Strike Activity," *American Economic Review*, March 1969, *59*, 35–49.

Christenson, C. Lawrence, "The Theory of the Offset Factor: The Impact of Labor Disputes upon Coal Production" *American Economic Review*, September 1953, *43*, 513–47.

Cragg, John G. and Uhler, Russell, S., "The Demand for Automobiles," *Canadian Journal of Economics*, August 1970, *3*, 386–406.

Cramton, Peter, "Bargaining with Incomplete Information: An Infinite-Horizon Model with Continuous Uncertainty," Research Paper No. 680, Graduate School of Business, Stanford University, May 1982.

Farber, Henry S., "Bargaining Theory, Wage Outcomes, and the Occurrence of Strikes," *American Economic Review*, June 1978, *68*, 262–71.

Freeman, Richard B. and Medoff, James L., "New Estimates of Private Sector Unionism in the United States," *Industrial and Labor Relations Review*, January 1979, *32*, 143–74.

Fudenberg, Drew, Levine, David and Ruud, Paul, "Strike Activity and Wage Settlement," Working Paper No. 249, UCLA, January 1983.

______ **and Triole, Jean,** "Sequential Bargaining with Incomplete Information," *Review of Economic Studies*, November 1981, *50*, 221–48.

Grubert, Harold, "An Empirical Study of the Economics of Bargaining," unpublished doctoral dissertation, MIT, 1968.

Harrison, Alan and Stewart, Mark, "Cyclical Variation in Individual Conditional Strike-Settlement Probabilities," Working Paper No. 85–20, McMaster University, July 1985.

Hayes, Beth, "Unions and Strikes with Asymmetric Information," *Journal of Labor Economics*, January 1984, *2*, 57–83.

Kennan, John, "The Duration of Contract Strikes in U.S. Manufacturing," *Journal of Econometrics*, forthcoming 1986.

Lewis, H. Gregg, "Union Relative Wage Effects: A Survey of Macro Estimates," *Journal of Labor Economics*, January 1983, *1*, 1–27.

Mauro, Martin J., "Strikes as a Result of Imperfect Information," *Industrial and Labor Relations Review*, July 1982, *35*, 522–38.

Reder, Melvin and Neumann, George R., "Conflict and Contract: The Case of Strikes," *Journal of Political Economy*, October 1980, *88*, 867–86.

Sobel, Joel and Takahashi, Ichiro, "A Multistage Model of Bargaining," *Review of Economic Studies*, July 1983, *50*, 411–26.

Tracy, Joseph, "A Theoretical and Empirical Investigation of U.S. Strike Activity," unpublished doctoral dissertation, University of Chicago, 1984.

U.S. Department of Labor, Bureau of Labor Statistics, *Current Wage Developments*, various issues.

______, *Industrial Relation Facts*, various issues.

______, *Wage Calendar*, various issues.

Marriage and Divorce: Informational Constraints and Private Contracting

By H. ELIZABETH PETERS*

Accompanying the striking rise in U.S. divorce rates during the last several decades have been changes in the laws regulating divorce. One of the more pervasive changes is the move towards "no-fault" divorce. Since 1970, more than half of the states have adopted a rule which basically allows divorce at the demand of either spouse.[1] The remaining states require either mutual agreement (explicitly or in practice) or—in the case of a contested divorce—a costly legal battle. The constraints imposed by the law and any shift in the law could have behavioral and distributional consequences. The rules might influence an individual's decisions about whether or not to divorce or to marry, and how much to invest in the marriage relationship. The law's impact on the nature of the marriage agreement might also affect the distribution of income within marriage and the terms of the divorce settlement.

This paper is an empirical study that utilizes a contract-theoretic framework to examine the impact of both legal and informational constraints on several aspects of the marriage relationship: 1) the probability of divorce; 2) compensation at divorce (i.e., the terms of the divorce settlement); 3) the probability of entering marriage; and 4) incentives for investment in marriage-specific capital. Two models of contracting that differ most fundamentally in their assumptions about information are contrasted. The first assumes that *ex post* information about the value of each spouse's opportunities at divorce is symmetric. The model predicts that costless bargaining could redistribute the gains to marriage in such a way that divorce only occurs when the joint benefits exceed the joint costs. Thus the law would have no effect on the divorce rate. The compensation scheme would, however, vary with the law to achieve efficiency. This argument is a form of the Coase Theorem. A second model asserts that the existence of asymmetric information might lead to a fixed wage marriage contract. This model predicts that the divorce rate would be higher in states that allow unilateral divorce.

The empirical investigation is based on data from a special 1979 *Current Population Survey* that contains information on marital history and financial settlements at divorce. Evidence that divorce is no more likely to occur for women living in unilateral divorce states, but that divorce settlement payments are lower for those women, supports the symmetric information hypothesis.

I. Marriage Contracts

This section presents two models of marital contracting. For each model, empirical implications are derived that specify the effects of different divorce laws (one law allows either spouse to initiate divorce *unilaterally*, the other law requires *mutual consent* before

*Assistant Professor, Department of Economics, and Research Associate, Research Program on Population Processes, Institute of Behavioral Science, University of Colorado, Boulder, CO 80309. I thank Gary Becker, Charles Kahn, Edward Lazear, Dorothy Maddi, Robert Michael, Donald Parsons, Sherwin Rosen, Lawrence Summers, Donald Waldman, and Franklin Zimring for helpful comments and encouragement. The paper has also benefited from comments by Orley Ashenfelter, two anonymous referees, and seminar participants at the University of Chicago, and Yale, Columbia, Michigan State, and Ohio State universities. Research support from the Economics Research Center/NORC, the Earhart Foundation, and the U.S. Department of Labor, Employment and Training Administration is gratefully acknowledged. I remain solely responsible for the paper's content.

[1]By this date, almost every state has adopted some form of no-fault divorce. However, only some of these no-fault laws allow for unilateral divorce. The classification scheme that I utilize is discussed later in the paper. See Doris Freed and Henry Foster (1979) for more details concerning the laws.

divorce can occur) on the probability of marriage and divorce and on the structure of the compensation scheme in the marriage contract. The first model assumes that information about the value of opportunities that would exist should divorce occur is symmetric, that is, the husband and wife have the same information. This model predicts that marriage contracts will be efficient: divorce only occurs when the joint value of the marriage is less than the sum of the values of opportunities that face each spouse at divorce. Thus, the probability of divorce will be independent of the divorce law, but the compensation scheme will vary with the law to induce efficient separation decisions. These results are a form of the Coase Theorem.

There are several reasons why this model might fail. One reason refers to the problem of imperfect state verification. Each spouse may not be able to determine the value of opportunities at divorce that face the other spouse. Incentives then exist for each spouse to misrepresent the value of alternatives that he or she faces and engage in costly strategic bargaining. The second model illustrates one possible solution to this asymmetric information problem: a fixed marriage wage contract. The contrasting empirical implications of this kind of contract are presented later.

The symmetric information model might also fail because of the problem of moral hazard. It may be difficult to determine each spouse's inputs to the marriage and to pay each the full return on his or her investment. Thus there exist incentives to invest in less than the optimal amount of marriage-specific capital. These incentives can be affected by the different constraints imposed by the divorce law.

The general structure and notation common to these models of marriage contracts is outlined below. Marriage is viewed as a long-term match between two individuals that produces a valuable, though partially intangible, output. Examples of this output include children, love, security, companionship, money income from market work, and household goods from home production.[2] Decisions about marriage and divorce are made as follows. Each individual who contemplates marriage has certain implicit expectations about the value of the marriage, the composition of the output, and the inputs required in the production of that output. That individual also has some idea about the probability of divorce, and the costs and consequences associated with the divorce. If the expected gains to marriage exceed the expected losses due to the possibility of future divorce, the marriage occurs. As the marriage continues over time, information about the value of alternatives, should divorce occur, becomes available. Examples of alternatives include the value of a potential new relationship (i.e., remarriage after divorce), and the value of market opportunities which might be different at divorce for women who were homemakers during marriage. Given this knowledge, the couple then decides whether to continue the marriage or to divorce.

To simplify the model, assume there are only two periods. At the beginning of the first period, the contract is negotiated. If it is acceptable to both, the couple marries and each spouse invests in the specified amount of marriage-specific capital. The information that is available at the time the contract is negotiated includes the present value to each individual of remaining single (S_w for the female and S_h for the male),[3] the joint probability *distribution* of possible divorce alternatives facing each spouse should marriage occur but later end in divorce ($g(A_w, A_h)$), and the joint value of marriage in each period (M_1 and M_2). At the beginning of the second period, the actual values of each spouse's opportunities outside the marriage are realized (A_w and A_h for the wife and husband, respectively), and a decision is made about whether to continue the marriage or to divorce. The expected joint value of the contract, R, can be expressed as the sum of the values of the possible outcomes in the two periods where each outcome is weighted by

[2] See Gary Becker (1974) for the development of this idea.

[3] In a more general model with more than two periods, the value of being single also includes the value of continuing to search for a marriage match.

the probability of its occurrence:

$$(1)\quad R = M_1 + b\left[M_2(1-p) + E(A_w + A_h | divorce)\cdot p\right],$$

where p is the probability of divorce, E is the expectation operator, and b is a discount factor. In this framework the stochastic variables are A_w, A_h, and p, where p is a function of A_w and A_h. The variables M_1 and M_2 are assumed to be fixed and known with certainty at the time the couple decides to marry. This last assumption implies that the inputs or marriage-specific capital investments of the husband and wife are fixed. Thus, the possibility of moral hazard is ignored in this specification. Later, the implications of relaxing this assumption are discussed. Because the true values of A_w and A_h are not revealed until period two, decisions about divorce are independent of M_1.[4] In the subsequent analysis of divorce rates and compensation, I ignore the first period and drop the discount factor and the subscript on M.

The inequality $R > S_w + S_h$ implies that there is a net gain to entering marriage, and some allocation of that gain is possible which would make both parties agree to marry. The *ex ante* division of the total product of the contract is determined in the marriage market. It will depend on such variables as the value to each of remaining single (i.e., the reservation wage), and the expected contribution (i.e., marginal product) and investment of each spouse in the marriage. The allocation of total compensation between a wage during marriage and compensation at divorce (possibly zero) depends on the probability of divorce, the distribution of alternatives available to each spouse at divorce, attitudes toward risk, and the legal constraints imposed on the contract.[5]

One assumption in the results derived below is that M, A_w, and A_h are perfectly divisible, that is, convertible into monetary units. Therefore the distribution to the husband and wife of the gains to marriage or divorce will fully exhaust those gains. This may not be an entirely plausible assumption in the analysis of marriage contracts because children are a major component of the gains to marriage, and they also play an important role in any divorce settlement.[6] One solution would be to posit a more general utility function that includes as separate arguments money and nonmonetary goods such as children. Incorporation of these issues in the present analysis, however, is beyond the scope of this paper.

In the following, I use this basic contract framework, which integrates the various aspects of the marriage relationship, to derive empirically testable implications for the two models.

A. *Symmetric Information*

1. Divorce and Compensation. Efficiency requires that resources be used in their highest valued activity. Because a husband and wife work together to produce the marriage output, joint wealth is maximized by a rule which compares the sum of each spouse's alternatives at divorce with the joint product of the marriage. Thus divorce in period two is efficient when

$$(2)\qquad M < A_w + A_h.$$

Figure 1 represents the joint distribution of all potential A_w and A_h.[7] The wife's opportunities at divorce get better the further from the origin is the realized value of A_w, whereas

[4] This result is crucially dependent on the assumption of no moral hazard problems.

[5] The nature of this contractual relationship is partly determined from the body of family law. Husbands have certain legal obligations towards their wives and vice versa. Lenore Weitzman (1980) examines the marriage contract implied by these laws. In this paper I focus on the particular aspect of the law that specifies which party can initiate divorce.

[6] Robert Mnookin and Lewis Kornhauser (1979) speculate about the possibility of a tradeoff between financial settlements at divorce and child custody arrangements.

[7] I have modified a figure which is presented in Masanori Hashimoto and Ben Yu (1980) and Robert Hall and Edward Lazear (1984) to fit this analysis of marriage.

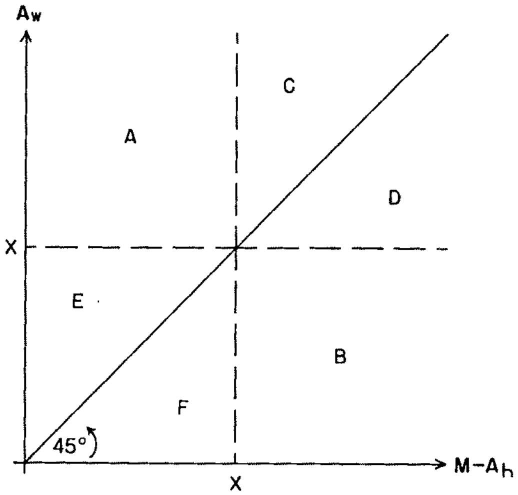


FIGURE 1. DISTRIBUTION OF ALTERNATIVES AT DIVORCE RELATIVE TO THE VALUE OF MARRIAGE

the husband has better divorce opportunities the closer $M - A_h$ is to the origin. Divorce is optimal anywhere to the left of the 45° line (areas A, C, and E) where $M - A_h < A_w$.

Now assume that the outcome of the contracting process specifies an initial marriage wage, X, for the wife and the residual, $M - X$ (which is fixed because by assumption M is fixed), for the husband. The wife then wants to divorce in period two when her realization of A_w is greater than X (areas A, C, and D). The husband wants to divorce when $A_h > M - X$ or $M - A_h < X$ (areas A, E, and F). The two agree to divorce in area A but disagree about divorce elsewhere (they both, however, want to stay married in area B). If the contract can be costlessly renegotiated in period two, *and* if each spouse has the same information about A_w and A_h, a compensation scheme exists such that only joint wealth-maximizing separations will occur.[8] The nature of the scheme depends on the rules under which divorce can be obtained.

If the law requires mutual consent, some kind of severance pay or compensation at divorce will be necessary to convince the wife

TABLE 1—INCOME UNDER DIFFERENT LEGAL REGIMES

Period Two Income	Mutual	Unilateral
Wife's		
If Married	X	$A_w + \alpha(M - (A_w + A_h))$
If Divorced	$A_w + C_w = X + \alpha(A_w + A_h - M)$	A_w
Husband's		
If Married	$M - X$	$A_h + (1-\alpha)(M - (A_w + A_h))$
If Divorced	$A_h + C_h = (M - X) + \alpha(A_w + A_h - M)$	A_h

(area E) or the husband (area C) to agree to a divorce. A divorce compensation formula like the following would induce efficient divorce:

$$(3)\quad C_w = (X - A_w) + \alpha(A_w + A_h - M).$$

$$(4)\quad C_h = -C_w = ((M - X) - A_h) + (1-\alpha)(A_w + A_h - M),$$

where C_w is what the wife receives (or pays out) and C_h is what the husband pays out (or receives). Then income at divorce is $A_w + C_w$ for the wife and $A_h + C_h$ for the husband (see Table 1). The first term in the compensation formulas is required to make each spouse indifferent between divorce and continuing the marriage. The second term represents the net gain to divorce, and α, the wife's share of that gain, is determined in the *ex ante* negotiations of the contract. Only when the second term is greater than zero will both husband and wife agree to divorce. Thus, efficiency and the requirement of mutual consent imposed by the law will jointly be satisfied.

If the law allows unilateral divorce, then either spouse can walk out of the marriage *without compensating the other*. In essence, both must agree for the marriage to continue. A *marriage* compensation scheme like the following would induce the wife (area D) or the husband (area F) to remain married. The wife receives

$$(5)\qquad X = A_w + \alpha(M - (A_w + A_h))$$

[8]See Becker et al. (1977).

and the husband receives the residual

$$(6)\ M-X=A_h+(1-\alpha)(M-(A_w+A_h)).$$

Analogous to the above discussion, the first term is the payment that makes each spouse indifferent between continuing the marriage and divorcing. The second term is the net gain from the marriage. If that gain is negative, then, as efficiency dictates, at least one spouse will decide to become divorced.

Under both laws divorce occurs when $A_w + A_h > M$. Thus the constraints imposed by the legal rules about the right to divorce do not affect the divorce rate. By contrast, the actual wage payments or income streams in the married and divorced states do depend on the divorce rule.[9] The compensation (in marriage or divorce) fluctuates so as to maximize the *ex post* value of the contract.

The relationship between the variance in divorce compensation payments under different legal regimes is straightforward. In this simple model, a mutual divorce law requires that C_w is paid before divorce can occur, and the variance of C_w is

$$(7)\quad \mathrm{Var}(C_w)=(\alpha-1)^2\mathrm{Var}(A_w|Z) + \alpha^2\mathrm{Var}(A_h|Z)+\alpha(\alpha-1)\mathrm{Cov}(A_w,A_h|Z),$$

where Z is the condition that $A_w + A_h > M$. In unilateral divorce states, no divorce compensation payments are required: $C_w = 0$ and $\mathrm{Var}(C_w) = 0$.

In a more complicated model there may be other reasons for divorce compensation payments to exist in unilateral as well as in mutual states. One reason is compensation for observable marriage-specific investments. A second reason relates to the division of tangible marital capital that is left after the marriage dissolves. As long as divorce compensation payments due to these other reasons are uncorrelated both with the unilateral/mutual divorce law and with A_w and A_h, then it still is true that the variance in the total divorce compensation payments under a mutual law is greater than the variance in those payments under a unilateral law.

The relationship between the variance in income at divorce under different legal regimes is more complex. Under a mutual divorce law regime, the first term of C_w guarantees that the wife will be no worse off at divorce than she was in marriage (see Table 1). This contract thus provides insurance against the possibility of receiving low values of A_w at divorce. The second term of C_w allocates the gains to divorce. Under a unilateral divorce law regime, income is just A_w.

Because of the insurance component of C_w in mutual divorce states, it seems intuitive that incomes at divorce should be less variable under mutual law than under unilateral law. Because of the variation caused by the second term of C_w, however, this intuition only holds under certain conditions. The variance in income at divorce in mutual law states is defined as follows:

$$(8)\quad \mathrm{Var}(A_w+C_w|Z)=\alpha^2\left[\mathrm{Var}(A_w|Z) + \mathrm{Var}(A_h|Z)+2\,\mathrm{Cov}(A_w,A_h|Z)\right].$$

Assume that the gains to divorce are split equally (i.e., $\alpha=1/2$) and that A_w and A_h have a bivariate normal distribution with $\rho=0$.[10] Then $\mathrm{Var}(A_w+C_w|Z)<\mathrm{Var}(A_w|Z)$

[9] If the law allows unilateral separation but also will enforce a compensation scheme where there is a separation penalty such as specified in equations (3) and (4), then individuals will behave as if mutual agreement is required, and the results for unilateral and mutual states will not be empirically differentiable. Thus it is important to examine the effect of the law on both the divorce rate and the compensation at divorce.

[10] Zero correlation between A_w and A_h is a convenient and not unreasonable assumption. The common idea in the marriage literature of positive assortative mating leads to a correlation between the traits that were observable when the couple married. There is no compelling reason, however, to assume that the random values of A_w and A_h would be correlated, because these values were not known when the couple married.

if $\mathrm{Var}(A_h) < 3\,\mathrm{Var}(A_w)$.[11,12] The intuition that the variance in divorce income under mutual law is less than the variance in divorce income under unilateral law holds as long as the relative variances of A_w and A_h are not too different.

2. Marriage. As seen in equation (1), the likelihood of entering marriage depends on the *ex ante* value of the contract, which, in turn, is a function of the probability of divorce and the value of alternatives at divorce. In the symmetric information model, the probability of divorce and $E(A_w + A_h|\text{divorce})$ do not depend on the divorce law. Therefore the initial constraints to the contract imposed by these different divorce laws will not affect the likelihood of entering marriage. The possibility of risk aversion, however, could induce differences in the propensity to marry because the variance in the value of the contract is a function of the divorce rule.

B. *Asymmetric Information*

This section focuses on contracting problems caused by bilateral asymmetric information. If neither spouse knows the value of the divorce opportunity facing the other, a rule relating compensation to these outside opportunities may not be viable. Each spouse would have an incentive to misrepresent his own opportunities or to participate in excess search for alternatives to the current marriage in order to gain a larger share of the marital wealth.[13] Behavior to support these claims can be costly and unproductive.

Because these incentives can be anticipated at the time the contract is made, an alternative to the *ex post* compensation scheme may be devised. One solution is a fixed wage contract which prohibits *ex post* bargaining.[14,15] The cost of this solution is the possibility of inefficient separations.[16]

1. Divorce and Compensation. Implications from this model for the divorce rate can be obtained from Figure 1. If the wife's marriage wage is fixed at X and bargaining is successfully prevented, a mutual divorce law will allow divorce only in area A where both agree. Since divorce is efficient in areas $A + C + E$, the mutual rule allows *fewer* divorces than is optimal. A unilateral rule leads to divorce in areas $A + C + E + F + D$ where either spouse wants to divorce. This rule produces *more* divorces than is optimal. Thus the fixed wage model would predict a higher divorce rate in unilateral states than in mu-

[11] Given the assumptions in the text, the relationship between the conditional and unconditional variance can be calculated as follows:

$$\alpha^2[\mathrm{Var}(A_w + A_h|Z)] = \alpha^2[\mathrm{Var}(A_w)+\mathrm{Var}(A_h)]\cdot Q_1$$

and $\mathrm{Var}(A_w|Z) = \mathrm{Var}(A_w)\cdot Q_2$, where $Q_1 = 1 - q(a)(q(a)-a)$; $Q_2 = 1-\rho_T^2 q(a)(q(a)-a)$; $a = M\cdot(\mathrm{Var}(A_w + A_h))^{-.5}$; $q(a)$ is the inverse Mills ratio evaluated at a; and ρ_T is the correlation between A_w and the truncated variable $A_w + A_h$ (see Jagdish Patel and Campbell Read, 1982, p. 320). It is obvious that $Q_1 < Q_2$. Therefore the sufficient condition in the text will be a lower bound on the variance of A_h relative to A_w.

[12] Given the same assumptions, a similar condition holds for the variance in husband's income at divorce: $\mathrm{Var}(A_h + C_h|Z) < \mathrm{Var}(A_h|Z)$ if $\mathrm{Var}(A_w) < 3\mathrm{Var}(A_h)$.

[13] Dale Mortensen (1978) discusses a related problem of excess job search in response to an offer matching scheme when search effort cannot be verified.

[14] Becker (1964) discusses the implications of a fixed wage contract for the incidence of quits and layoffs when the worker-firm match involves specific capital. More recently, Hashimoto and Yu and Hall and Lazear explore various motivations for fixed wage contracts.

[15] An important consideration in determining the feasability of fixed wage contracts is whether the no *ex post* bargaining stipulation can be enforced. As long as there is a positive benefit to strategic *ex post* bargaining, an incentive for that behavior exists. The multiperiod aspect of the marriage relationship may, however, reduce incentives for strategic bargaining. The mistrust and ill-feelings engendered by strategic bargaining can be detrimental to a relationship based on trust and intimacy. The greater the reduction in the value of marriage in future periods, the less bargaining that will occur.

[16] The fixed wage contract is not the only solution that addresses the problems caused by asymmetric information. More complicated game theoretic schemes have been proposed which can lead to optimal separations (see Michael Riordan, 1984). These schemes do not appear to be reflected in real world marriage contracts, however. Therefore these first-best contracts will be ignored in the subsequent analysis.

tual states. Because the amount of efficient divorce is independent of X (i.e., it occurs when $M < A_w + A_h$), this result will hold for any value of X.[17]

The determination of X is based on *ex ante* information,[18] and is calculated to minimize the potential inefficiencies—areas C and E where divorce should occur but does not under the mutual rule, or areas D and F where marriage should continue but does not under the unilateral rule. In this model the marriage wage plays the same role as the optimal sharing of investment costs in Gary Becker's (1964) model of specific human capital investment. The sharing rule (i.e., the compensation scheme) depends in part on the distribution of each party's alternatives to the match (for example, the relative variances and covariances) and on the separation rule imposed by the law (see Masanori Hashimoto and Ben Yu; Robert Hall and Edward Lazear).

2. Marriage. Joint incentives to marry depend on the expected value of the contract. The existence of *ex ante* rents to marriage is important because the divorce rate differs under the two laws. If $E(M) > E(A_w + A_h)$, the expected value of the marriage contract will be higher under the law that minimizes the number of inefficient divorces. Thus, more couples may be expected to marry in a mutual law state.

C. *Moral Hazard*

The previous analysis dealt primarily with the issue of the availability of information about *alternatives*. When the effort or other inputs of each party can affect the value of the relationship, however, information about the *inputs* is also necessary to avoid the problems associated with moral hazard. In the labor market literature, it has sometimes been assumed that it is easier to verify the value of alternatives (for example, an alternative wage offer) than to observe the amount of effort on the job. This same kind of problem can exist in marital relationships as well.[19]

A woman who invests in marriage by working in the household may face a lower market wage in the future. Because she only receives a fraction of the value of her investment in marriage, but bears the full cost of the lower wages at divorce, she has an incentive to self-insure against these potentially negative consequences by working more in the labor market during marriage. Marital fertility is another kind of behavior that can alter both the value of marriage and the value of alternatives at divorce. Divorced women with children have a higher cost of entering the labor market and also are less likely to remarry. Women who foresee the possibility of divorce might decide to have fewer children than they would otherwise.

To the extent that this behavior can be anticipated or observed, the structure of the initial contract will be altered to provide incentives which mitigate the problem. Constraints imposed by the law that limit the contracting responses reduce the value of the relationship and increase the likelihood of this kind of strategic behavior.

Divorce settlement payments are one way to compensate the wife for her previous marital investments. However, when the law allows either spouse to dissolve the marriage unilaterally, it may be more difficult to enforce this kind of payment. The unilateral divorce rule might therefore lead to an increase in the divorce rate as the value of marriage falls relative to the value of alternatives at divorce. Conversely, if the potential

[17] The value of X will influence the magnitude of the difference in divorce rates between the two legal regimes because it influences the relative amount of efficient vs. inefficient decisions. But the direction of the difference is independent of X.

[18] It is possible to have partially contingent payments that are based on jointly observable *ex post* information such as age and education of each spouse, number of children, etc.. Thus the contract could vary among observationally distinct couples. In this case, A_w and A_h might be viewed as the residual divorce alternatives obtained after controlling for observable characteristics.

[19] See Elisabeth Landes (1978) and William Johnson and Jonathan Skinner (1986) for more formal treatments of moral hazard problems in marital relationships.

TABLE 2—VARIABLE DEFINITIONS

Variable	Definition
Ever Divorced	= 1 if the woman became divorced during 1975–78.
Remarried	= 1 if a divorced woman was remarried by April 1979.
Time Since Divorce:	
< 1 year	= 1 if divorce occurred less than 1 year before April 1979.
< 2 years	= 1 if divorce occurred 1 to 2 years before April 1979.
< 3 years	= 1 if divorce occurred 2 to 3 years before April 1979.
< 4 years	= 1 if divorce occurred 3 to 4 years before April 1979.
Age	= age of woman in 1979.
Age at Divorce	= age at divorce for women who became divorced.
Education	= number of years of education completed as of 1979.
White	= 1 if woman is white.
Kids under 18	= number of children younger than 18 in the household in 1979; for ever-divorced women this number only includes children from the former marriage.
Kids Squared	= Kids under 18 × Kids under 18.
Kids under 6	= number of children younger than 6 in the household in 1979; the number is not calculated for remarried women because it is difficult to separate children from the former marriage.
Kids 6–18	= number of children ages 6 to 18 in the household; the same restriction applies as for kids under 6.
SMSA	= 1 if the woman lived in a standard metropolitan statistical area in 1979.
In Labor Force 1978	= 1 if the woman had positive earnings in 1978.
Husband's Earnings	= husband's earnings in 1978 in 1000's of dollars.
Earnings	= woman's earnings in 1978 in 1000's of dollars.
Family Income	= total family income in 1978 in 1000's of dollars.
Unilateral	= 1 if the woman lived in a unilateral state at the time of divorce (or remarriage).
State Divorce Rate 1970	= number of divorces per 100 women at risk of divorce in 1970 in the state in which the woman lives.
State Catholic 1970	= percent Catholic in 1970 in the state in which the woman lives.
Alimony Received	= amount of alimony received in 1978.
Child Support Received	= amount of child support received in 1978.
Support	= Alimony Received + Child Support Received
Settlement Value	= value of the property settlement received.
Proportion Eligible	= proportion of 1978 a woman was eligible to receive payments; if divorced or remarried during 1978, Proportion Eligible is the time after the divorce or before remarriage; otherwise Proportion Eligible = 1.

for moral hazard can be anticipated, only those marriages with expected returns that are high enough to offset these costs will occur. The impact of the unilateral rule would then be seen as a fall in the rate of marriage rather than as an increase in the rate of divorce.

D. *Summary of Theoretical Results*

The most clearly distinguishing implication of the two contract models concerns the divorce rate. If information about alternatives for each spouse at divorce is symmetric, a contingent compensation scheme can be devised such that divorce only occurs when it is *ex post* efficient. Thus the divorce rate will be the same regardless of the constraints imposed by law. The compensation scheme, however, would vary with the law to enable this efficient outcome. If a problem of asymmetric information leads to the development of fixed wage contracts, the divorce rate would be higher when unilateral dissolution is allowed.

The impact of the law on other aspects of the marital relationship—for example, the propensity to marry and the compensation at divorce—is less easy to predict. Assumptions about the distribution of divorce opportunities for the husband and wife, and the sharing ratio, α, of the rents from marriage and divorce, as well as the extent of asymmetric information about the *ex post* divorce alternatives are crucial. The issue of whether either spouse can influence the value of the

TABLE 3—MEANS FROM 1979 *Current Population Survey* DATA

Variable	All		Ever Divorced		Remarried	
	Unilateral	Mutual	Unilateral	Mutual	Unilateral	Mutual
Ever Divorced (percent)	6.2[a]	5.3[a]	100.	100.	100.	100.
Remarried (percent)	–	–	31.6	33.2	100.	100.
Age	43.3[a]	44.0[a]	35.5	36.1	33.9	34.2
Education	12.9[a]	12.8[a]	13.1	12.9	13.0	12.7
White (percent)	91.7[a]	89.3[a]	90.3	89.3	92.9	94.1
Kids under 18	1.0[a]	1.0[a]	1.0	0.9	0.8[a]	0.6[a]
SMSA (percent)	50.9[a]	58.8[a]	54.7	57.8	53.6	51.2
In Labor Force 1978 (percent)	56.6[a]	54.0[a]	82.4[a]	77.5[a]	75.0[b]	67.1[b]
Earnings	3,626.	3,594.	6,407.	5,974.	5,004.	4,427.
Family Income	20,731.	20,815.	14,242.	14,015.	22,670.	21,213.
Alimony Received[c]	–	–	200.[a]	505.[a]	–	–
Child Support Received[c]	–	–	1,070.[a]	1,472.[a]	–	–
Settlement Value[c]	–	–	6,752.	6,605.	–	–
Sample Size	11,500	9,714	709	512	224	170

Note: Those excluded from the analysis are women who had moved from another state since 1975 and either were married or became divorced in 1975–78. A second group of women who were excluded from the analysis includes never married women, widows, and those women who became divorced before 1975.

[a] Mutual and unilateral means are not equal at the 5 percent significance level.

[b] Mutual and unilateral means are not equal at the 10 percent significance level.

[c] Only 7.2 percent of the sample of ever-divorced women received any alimony in 1978. For those with positive amounts, the mean alimony payment was $3456. In 1978, 62.6 percent of eligible women received child support and 58.1 percent received a property settlement at the time of divorce. The mean values for those with positive amounts were $2145 and $11,522 for child support and property settlements, respectively.

marriage as a means of gaining an advantage at divorce may also be important.

II. Empirical Analysis

The data consist of observations from the March/April 1979 *Current Population Survey* that includes demographic variables such as state of residence, age, current marital status, number of children, and economic variables such as income, labor force behavior, and education. Additional information on marital history, type and amount of child and spousal support payments, and the value of any property settlement is available for those women who have ever been divorced.

To study the division of family resources at divorce and the probability of remarriage (used to investigate the implications for marriage in general), I use the sample of 1,221 women who became divorced between 1975 and 1978, and who indicated that they had lived in the same state in 1975 as in 1979. These requirements are so that I can make some inference about the state in which a woman became divorced.[20] For estimates of the divorce rate, this ever-divorced sample is combined with the sample of currently married, never-divorced women who lived in the same state in 1975 as in 1979. Table 2 gives the variable definitions and Table 3 presents the means for the variables.

Information on state divorce laws during 1975–79 is also available. States are grouped according to whether they fit in the following categories (Table 4 lists the states in each):

1) Pure unilateral—either spouse can dissolve the match by saying there are "irreconcilable differences," or that the marriage is "irretrievably broken." This is commonly

[20] The possibility exists that a woman may obtain a divorce in a state other than the one in which she resides. The problem of "migratory divorce" is briefly examined in Hugh Carter and Paul Glick (1976). Their estimates of out-of-state divorces range from one-in-ten to one-in-twenty in the period before the easing of divorce laws in the early 1970's. Presumably migratory divorce would be much less frequent today.

TABLE 4—CLASSIFICATION OF STATES BY 1978 DIVORCE LAW

	Unilateral[a]		
Arizona	Michigan	New Hampshire	Alaska
California	Montana	New Mexico	Georgia
Colorado	Nebraska	Indiana	Oklahoma
Florida	Oregon	Connecticut	Kansas
Iowa	Washington	North Dakota	Hawaii
Kentucky	Wyoming	Maine	Idaho
Minnesota	Nevada	Rhode Island	Texas
Alabama	Massachusetts		
	Mutual[b]		
Illinois	Ohio	West Virginia	Missouri
Pennsylvania	Tennessee	New Jersey	Virginia
South Dakota	Utah	Louisiana	Arkansas
North Carolina	Mississippi	Wisconsin	Maryland
New York	South Carolina	Vermont	Washington, D.C.
Delaware			

Source: Freed and Foster (1979).

[a] Includes states with only irretrievable breakdown and incompatability grounds for divorce and states where these grounds have been added to traditional fault grounds.

[b] Includes states with only fault grounds for divorce, states that explicitly stipulate mutual agreement, and states where divorce is not allowed until the couple has been separated for a specified period of time.

called no-fault. (Included in this group are states which allow unilateral no-fault divorce but have still retained the traditional fault grounds. In practice, however, in these states the unilateral rule dominates.)

2) Mutual divorce with binding arbitration should no agreement be reached. What usually happens in "fault only" states is that if the couple agrees to divorce there is collusion. One spouse claims the other is guilty of one of the traditional grounds (usually cruelty) and no defense is made by the other spouse. If they cannot agree and one spouse sues for divorce, the court decides whether or not to grant the divorce. A second type of mutual divorce occurs in states which use no traditional fault grounds but require that the couple reach an agreement about the division of marital resources before a divorce is granted. If no agreement is reached, the court may stipulate the division and grant the divorce.

3) Unilateral divorce is possible but requires a large cost. States which fit this category are those, for example, which allow divorce on the grounds of unilateral separation for periods ranging from one to seven years. If these costs are large enough it will be cheaper to agree than to divorce unilaterally. Thus in the empirical analysis I have included these states under the mutual divorce category.

A. *Divorce Rate*

The theory presented earlier predicted that if there were an effective constraint to the spouses' *ex post* bargaining about the individual returns to marriage or the division of resources at divorce, the divorce rate would be higher in states that allowed unilateral divorce. Popular wisdom agreed with that prediction. Much of the opposition to the legal change from mutual to unilateral divorce was based on the belief that if either spouse could walk out of the marriage, divorce would be easier and thus more likely to occur. A second model predicted that costless bargaining could redistribute the gains to marriage or divorce in such a way that divorce only occurred when the joint benefits exceeded the joint cost. Thus the law would have no effect on the divorce rate. In this section I present evidence about the relationship between the law and the divorce rate.

TABLE 5—PROBABILITY OF DIVORCE AND REMARRIAGE

	Divorce		
	(A)	(B)	Remarriage
Intercept	-3.44^{b} (5.15)	-2.81^{b} (4.91)	27.69^{b} (4.27)
Age[c]	-0.26^{b} (403.95)	-0.26^{b} (405.05)	-1.23^{b} (65.04)
Education	-0.19^{b} (8.57)	-0.19^{b} (8.95)	0.23 (0.13)
White	−0.38 (0.48)	−0.33 (0.37)	19.49^{b} (9.95)
Kids under 18	-1.53^{b} (20.80)	-1.55^{b} (21.33)	-7.49^{b} (5.16)
Kids Squared	0.07 (0.54)	0.07 (0.56)	0.70 (0.48)
SMSA	0.61^{a} (3.25)	0.56^{a} (2.87)	−0.25 (0.01)
South	1.05 (1.67)	1.81^{b} (12.47)	22.50^{b} (19.72)
West	1.70^{b} (4.57)	3.12^{b} (30.21)	20.46^{b} (13.50)
North Central	1.41^{b} (5.32)	1.91^{b} (13.86)	15.87^{b} (9.90)
In Labor Force 1978	–	–	-17.96^{b} (21.65)
Time Since Divorce:			
< 1 year	–	–	-41.65^{b} (38.08)
< 2 years	–	–	-26.15^{b} (21.73)
< 3 years	–	–	-16.35^{b} (9.13)
< 4 years	–	–	−4.36 (0.70)
State Divorce Rate 1970	0.96^{b} (17.36)	–	
State Catholic 1970	−0.23 (0.01)	–	
Unilateral	−0.32 (0.70)	0.01 (0.00)	-9.77^{b} (8.29)
Sample Size	21,214	21,214	1,153
Percent	5.76	5.76	28.27

Note: The estimates reported are $\partial \text{Percent}/\partial X = \beta(\bar{P}\cdot(1-\bar{P}))\cdot 100$ from the logit $P = 1/(1+e^{-X\beta})$. *Chi*-square statistics are in parentheses.
[a]Significantly different from zero at the 5 percent level.
[b]Significantly different from zero at the 10 percent level.
[c]In the remarriage regression this variable is age at divorce.

The dependent variable in these logit regressions is the probability of becoming divorced during 1975–78 given no previous divorces. Parameter estimates are reported in Table 5. As found in other studies (see, for example, Becker et al.), divorce is less likely to occur if a woman is older, more educated, has more children, and does not live in an urban area. The unilateral divorce rule is found to have no relationship to the probability that a woman becomes divorced.[21]

[21]To test whether this result is an artifact of the classification scheme, the same regression was run using a subsample of women living only in states with divorce

The state divorce rate in 1970 (a time period *before* the changes in divorce laws) is included as a proxy for an unobserved state-specific propensity to divorce. The coefficient is, as expected, positive and highly significant—the higher the 1970 state divorce rate, the more likely is a woman living in that state to become divorced during 1975–78. My earlier analysis (1983) found that states with higher 1970 divorce rates were more likely later to adopt a unilateral divorce rule. Thus leaving out the 1970 divorce rate would result in an upward bias on the estimated effect of unilateral divorce (compare regressions A and B in Table 5). The legal variable is, however, insignificant under either specification.

The *CPS* sample that I use is not representative of the population as a whole because it only includes women who lived in the same state in 1975 and 1979. Both theoretical and empirical research have found that geographic mobility is positively correlated with divorce (see Bill Fenelon, 1971; Jacob Mincer, 1978). Therefore it is important to determine whether this sample selection criterion seriously affects the conclusions drawn from Table 5 that the probability of divorce is independent of the law regulating divorce.

If migration is a *consequence* of divorce (for example, divorced individuals may tend to move to avoid bad memories and associations), and if, for some exogenous reason, interstate migration is higher for unilateral states, then omitting migrants from the analysis would cause a downward bias on the coefficient on unilateral divorce. This could explain why the results show that the divorce rate is not higher in unilateral states. Alternatively, if migration is a factor that *precipitates* divorce (for example, conflicting location choices for the husband and wife leading to marital breakdown), then omitting migrants from the sample would not alter the conclusion that, *ceteris paribus*, the probability of divorce is independent of the law regulating divorce.

Results from aggregate state analyses can be used to resolve this issue. Several univariate and multivariate time-series studies have examined changes in state aggregate divorce rates before and after the changes in the law that occurred during the 1970's (Gerald Wright and Dorothy Stetson, 1978; Harvey Sepler, 1981; Robert Schoen et al., 1975; Becker, 1981; and myself, 1983). For the most part, these studies found that the trend in the divorce rates was unaffected by the change in the law. A study of several Scandanavian countries that had implemented what were essentially unilateral divorce rules shows similar results (Annemette Sorensen, 1980). These studies, in conjunction with the results reported in Table 5, provide evidence in support of the symmetric information model.

B. *Compensation at Divorce*

The empirical test of the effect of the law depends crucially on the assumption that the actual practice in states with mutual divorce laws differs from that in states with unilateral divorce laws. For example, if the law allows unilateral separation but also enforces a scheme where the injured party is compensated for unanticipated loss at divorce, then individuals would behave as if mutual agreement were required. In this case it would be the compensation scheme rather than the separation rule which influences behavior. Thus it is important to examine jointly the law on the books and the compensation scheme which is practiced. This subsection explores the effect of the law on the levels of the divorce settlement and on the variability of the settlement.

The divorce settlement can consist of three kinds of payments: 1) alimony; 2) child support; and 3) property settlement. Although in theory the purpose for each payment is different, in practice they should be fungible. The weight given to each, however, will vary according to the importance of such factors as the differential tax structure applied to each type of payment, the fact that alimony

laws that could be unambiguously classified as unilateral or mutual. In these regressions, the probability of divorce is not higher in unilateral divorce states, and there is some evidence that the rate may be slightly lower.

TABLE 6—*OLS* REGRESSIONS OF AMOUNT OF DIVORCE SETTLEMENT RECEIVED

	Alimony Received	Child Support Received	Support Received	Settlement Value
Intercept	−1,423.22[b]	−4,265.92[b]	−3,323.69[b]	−21,614.50[b]
	(4.26)	(4.59)	(5.83)	(7.66)
Age of Divorce	22.20[b]	55.38[b]	39.63[b]	248.34[b]
	(5.77)	(4.27)	(6.04)	(7.37)
Education	59.47[b]	115.95[b]	113.91[b]	1056.90[b]
	(3.43)	(2.46)	(3.85)	(6.96)
White	247.72[a]	710.41[b]	689.89[b]	4317.20[b]
	(1.75)	(2.13)	(2.85)	(3.48)
Kids under 18	23.16	971.30[b]	1085.10[b]	3883.78[b]
	(0.27)	(3.35)	(7.38)	(5.14)
Kids Squared	−1.80	−138.87[b]	−154.23[b]	−607.07[b]
	(0.08)	(2.34)	(3.86)	(2.96)
SMSA	97.86	73.64	141.66	1642.55[b]
	(1.15)	(0.36)	(0.98)	(2.21)
In Labor Force 1978	−205.81[a]	241.30	−15.00	1284.10
	(1.86)	(0.88)	(0.08)	(1.32)
Proportion Eligible	321.45[a]	1225.39[b]	689.88[b]	–
	(1.71)	(2.33)	(2.15)	
Time Since Divorce	−65.19	−204.84	−120.85	−930.96[b]
	(1.41)	(1.65)	(1.53)	(2.95)
Unilateral	−185.65[b]	−462.36[b]	−448.54[b]	−137.19
	(2.19)	(2.25)	(3.10)	(0.18)
R^2	.048	.078	.105	.110
Sample Size	1,221	636[c]	1,221	1,221

Note: *T*-statistics are shown in parentheses.
[a]Significantly different from zero at the 5 percent level.
[b]Significantly different from zero at the 10 percent level.
[c]The regression is run on the sample of women with children who are eligible to receive child support.

and child support are usually periodic payments and the property settlement is lump sum, and the restriction that alimony ends at remarriage.

Table 6 presents the regression estimates. The regressions include several socioeconomic variables to control for the influence of the woman's observable characteristics on the divorce settlement. The estimates of these parameters are consistent with the expectation that compensation at divorce is positively related to the value of the marriage and to the amount of marriage-specific investment by the wife (see Landes for similar results concerning alimony).

Table 6 also shows that the amounts of alimony and child support received are significantly lower in unilateral states.[22,23] Evidence that the divorce rate does not differ between the two regimes but that the compensation schemes do differ supports the symmetric information model—separations are efficient, but the constraints imposed by the law require different compensation schemes to achieve that efficiency.

[22]The amount of the property settlement seems to be unrelated to the unilateral/mutual law. Thus one cannot make the argument that unilateral divorce was accompanied by other changes in marriage property law which tended to raise the property settlement and offset the lower levels of child support and alimony payments.

[23]A few studies have examined changes in divorce settlement payments in a single state before and after the adoption of a unilateral divorce rule (Ruth Dixon and Weitzman, 1980; K. Seal, 1983; Charles Welch and Sharon Price-Bonham, 1983). In general, these studies are consistent with the cross-state results reported in this paper. No-fault (or unilateral divorce) is associated with lower amounts of alimony and child support payments (although the differences were not always significant in the Welch and Price-Bonham study). Two of the studies also showed significant changes in property settlements.

TABLE 7—MEASURES OF THE DISPERSION OF THE DIVORCE SETTLEMENT AND FAMILY INCOME BY STATE LAW
(Shown in thousands of dollars)

		Variance	Coefficient of Variation[e]	Unexplained Variance[c]	Unexplained Coefficient of Variation[d,e]
Alimony Due	M[b]	6.28[a]	486.98	5.98[a]	475.43
	U	0.87	506.38	0.83	494.88
Alimony Received	M	4.11[a]	558.44	3.92[a]	545.39
	U	0.82	544.22	0.79	534.19
Child Support Due	M	14.60[a]	195.43	13.99[a]	191.29
	U	2.38	92.33	2.08	86.32
Child Support Received	M	13.54[a]	228.02	13.09[a]	224.21
	U	2.19	128.08	1.92	119.91
Support	M	12.42[a]	298.72	11.37[a]	285.87
	U	2.76	212.54	2.37	196.79
Settlement Value	M	204.14[a]	216.32	177.30[a]	201.59
	U	165.65	190.61	151.78	182.45
Family Income	M	124.45	77.53	114.70	73.95
	U	137.00	78.76	114.16	73.64

[a]Variances are not equal at 5 percent significance level.
[b]M = Mutual Divorce Law Sample; U = Unilateral Divorce Law Sample.
[c]Unexplained Variance is the sum of squared residuals divided by the degrees of freedom from a regression of the divorce settlement payment on individual characteristics.
[d]From previous column—see note c.
[e]Tests of significance are not calculated for coefficient of variation.

C. *Variance in Payments*

The variance in settlement payments in a symmetric information model is a function of the divorce law. Mutual divorce requires that these payments vary to compensate the spouse that is worse off at divorce. Compensating payments are not required under unilateral divorce. Thus the variance in settlement payments will be higher under a mutual divorce law if the information is symmetric.

To test for a difference in variances, a simple F-distribution is used on the ratio of the two variances. In all of the six measures of divorce payments in Table 7, the variance under mutual divorce is significantly greater than that under unilateral divorce. Because the distribution of observed characteristics may differ in states with the two laws, I also calculate F-tests on the residual variances from regressions of the settlement payments on the characteristics of divorced women. Again, variance is significantly higher in mutual divorce states. These tests are on absolute variances. The mean payment also differs by the divorce law, therefore a measure of relative variance—the coefficient of variation—is calculated. Using this measure reduces the differential in variances, but in only one case is the direction of that inequality reversed.

If divorce payments under a mutual law compensate for a low draw of A_w, under certain distributional assumptions, they should reduce the variance in total income. However, an F-test shows no significant difference in the variance of total income between unilateral and mutual states. The results from the F-tests on equality of variances of divorce payments tend to support the symmetric information model, and the results about variances in income are not inconsistent with the model.

D. *Remarriage*

The *CPS* data do not allow a test of the relationship between unilateral divorce law and first marriages. Some information, however, exists about remarriage of those previously divorced. The kind of contract made at

remarriage may differ in some significant ways from one made at first marriage (for example, the presence of children from a previous marriage may alter the nature of the agreement). I do not, however, expect the qualitative effect of the divorce rule on the contract to differ between marriage and remarriage. The effect may, in fact, be stronger because women who have already been divorced are more aware of any potential consequences of divorce laws.

The dependent variable in the last logit regression in Table 5 is the probability that a woman remarried by April 1979, given that she became divorced sometime between 1975 and 1978. The effects of the socioeconomic variables in this regression are consistent with other results (see Becker et al.). Older women and women with children are less likely to remarry. Whites are more likely to remarry. The date of divorce in this sample ranges from 1975 to 1978. The coefficients on the dummy variables indicating time since divorce are negative compared to the omitted category of divorce occurring more than 4 years before 1979. As expected, the probability of being remarried by April 1979 increases as time since divorce increases. Unilateral divorce lowers the probability of remarriage by a little more than nine percentage points.[24]

E. *Moral Hazard*

The largest source of income for divorced women who have not remarried is earnings. Staying in the home and caring for the children during marriage will lower future potential earnings. If this marriage-specific investment is not fully compensated for at divorce, there exists an incentive for married women to reduce their time spent in the home and increase their more general market capital by entering the labor force.

In Table 8, I examine the effect of unilateral divorce laws on the labor force participation of married women. The independent variables have the usual effect on labor force participation. Older women, white women, and women with young children are less likely to participate. The marginal effect of a child declines with the number of

TABLE 8—LABOR FORCE AND FERTILITY BEHAVIOR OF NEVER-DIVORCED MARRIED WOMEN

	Labor Force Participation[c]	Children Less than 6[d,e]
Intercept	71.98[b]	1.91[b]
	(426.38)	(25.73)
Age	−1.92[b]	−0.04[b]
	(2,119.43)	(22.81)
Kids under 6	−30.84[b]	–
	(752.80)	
Kids 6–18	−8.80[b]	–
	(99.30)	
Kids Squared	1.42[b]	–
	(54.94)	
Education	3.46[b]	−0.01[b]
	(452.74)	(3.23)
White	−12.52[b]	−0.11[b]
	(71.21)	(3.64)
Husband's Earnings (in thousands)	−0.39[b]	0.01[b]
	(98.21)	(5.52)
SMSA	−0.31	−0.03[b]
	(0.14)	(1.98)
South	−0.08	−0.02
	(0.00)	(0.80)
West	−0.71	0.06[b]
	(0.28)	(2.19)
North Central	2.54[b]	0.04[a]
	(4.98)	(1.74)
Unilateral	2.22[b]	−0.02
	(6.34)	(1.23)
Percent	53.64	–
R^2		0.06
Sample Size	19,501	8,413

[a]Significantly different from zero at the 5 percent level.

[b]Significantly different from zero at the 10 percent level.

[c]The estimates reported are ∂ Percent$/\partial X = \beta(\bar{P}\cdot(1-\bar{P}))\cdot 100$ from the logit $P = 1/(1+e^{-X\beta})$. *Chi*-square statistics are in parentheses.

[d]*T*-statistics for the *OLS* regression are in parentheses.

[e]The sample is restricted to women younger than 40.

[24]Because all individuals in this sample have not had the same amount of time in which to become remarried, the logit estimates of the probability of remarriage within a given time period will not be entirely correct. The use of hazard rate models might be more appropriate. However, the coefficient estimate on the variable of interest, the divorce law, is robust across logit specifications that both included and excluded the time since divorce control variables. Therefore, the limitations imposed by the specification in the paper are probably not too severe.

children. Husband's earnings are also negatively related to the likelihood of participation. Living in a unilateral state, however, increases the probability of participation by two percentage points. This may be indicative of the incentive for married women to enter the labor force to self-insure against becoming divorced without compensation.[25]

The analysis of fertility tells a somewhat different story. Table 8 reports a regression of the number of children younger than age 6 living in the households of never divorced, married women. The effect of the independent socioeconomic variables is consistent with other research. Fertility is lower for older, more educated, and white women, and is positively related to the earnings of the husband. The coefficient on the legal variable is negative, but not significantly different than zero.

Thus it seems that differences in law are associated with differences in labor force behavior, but not with differences in fertility behavior. There are several possible explanations for these results. First, the measure of fertility, the number of children younger than 6, represents fertility decisions made over a period in the past. It would take some time for changes that occurred in the legal environment during the mid-1970's to have their full impact on that measure. The labor force variable, however, measures a single year's participation decision. A response in labor force behavior to a change in the law would therefore be observed sooner.

A second explanation may involve the relationship between children and labor market opportunities and the amount of the divorce settlement for women. Better labor market opportunities raise the income for divorced women but do not consistently lower the divorce settlement (see Table 6). Children may increase the cost of being divorced; however, this is partly offset by the positive effect that children have on the divorce settlement. Thus a change in labor force behavior may be a more effective method for a woman to self-insure against the negative consequences at divorce.

It may seem surprising that this moral hazard problem does not lead to higher divorce rates in unilateral states. If, however, the problems can be anticipated, individuals may respond, instead, by marrying less or by searching longer for a better match. This may explain the results from the remarriage regressions.[26]

III. Conclusions

This paper presents empirical evidence about the impact of two types of divorce laws on various aspects of the marriage contract. One law specifies that either spouse can initiate divorce (unilateral), the other requires that both agree to divorce (mutual). Two models of contracting are proposed that differ in their assumptions about the information set available to each spouse. One model assumes that *ex post* information about the value of opportunities outside the marriage is symmetric. The other model assumes that information is asymmetric. The clearest distinction between the two models concerns predictions about divorce rates and divorce settlement payments. The empirical results show that the divorce rates are not significantly different in unilateral and mutual consent states, but that divorce settlement payments are lower in unilateral states. This evidence supports the hypothesis of the symmetric information model that divorce

[25] A second specification of this regression included the 1970 state married female labor force participation rate as an exogenous variable. While this variable was highly correlated with the dependent variable, its inclusion had a trivial impact on the coefficient for unilateral divorce. The participation rate for married women living in unilateral states in 1979 remained almost 2 percentage points higher. A similar kind of variable, 1970 state marital fertility rate was included in the fertility equation. Again, this specification did not affect the coefficient on unilateral divorce.

[26] Johnson and Skinner develop a model where women who anticipate a higher probability of divorce will invest more in the labor market to insure against lower earnings at divorce. The evidence presented in this paper suggests that it is the nature of the compensation scheme that is important. If a scheme could be enforced that insured women against the negative consequences of divorce, incentives for providing less than optimal investment in marriage would be reduced.

occurs when it is efficient, but the compensation scheme depends on the divorce law. Findings of lower remarriage rates and lower variance in compensation at divorce for women living in unilateral states are also consistent with this hypothesis.

Despite the recent emphasis in the contract literature on the constraints of asymmetric information (see the survey by Oliver Hart, 1983), there are several reasons why the above results may not seem unreasonable for the case of marriage. First, the nature of the interaction in marriage makes it more difficult to conceal information than would be true of more impersonal relationships. Second, there may be ties such as the presence of children that continue after the relationship has terminated. These ties may create incentives to reveal information by inducing each party to consider the joint, rather than the individual, gain to any action.[27]

[27]These results concerning marriage contracts could be relevant to certain labor market relationships with similar characteristics. These include jobs where there is a close working relationship between employer and employee, jobs with a high degree of firm-specific capital, and jobs which induce the individual to consider the profit of the company even after job separation through compensation such as stock options. Employees in a small firm and specialized executives and managers are some examples of the kinds of relationships to which the analysis and conclusions presented in this paper might be applicable.

REFERENCES

Becker, Gary S., *Human Capital*, NBER, New York: Columbia University Press, 1964.

______, "A Theory of Marriage: Part II," *Journal of Political Economy*, March/April 1974, *82*, 11–26.

______, *A Treatise on the Family*, Cambridge: Harvard University Press, 1981.

______, **Landes, Elisabeth M. and Michael, Robert T.,** "An Economic Analysis of Marital Instability," *Journal of Political Economy*, December 1977, *85*, 1141–88.

Carter, Hugh and Glick, Paul C., *Marriage and Divorce: A Social and Economic Study*, Cambridge: Harvard University Press, 1976.

Dixon, Ruth B. and Weitzman, Lenore J., "Evaluating the Impact of No-Fault Divorce in California," *Family Relations*, No. 3, 1980, *29*, 297–307.

Fenelon, Bill, "State Variations in United States Divorce Rates," *Journal of Marriage and the Family*, May 1971, *33*, 321–27.

Freed, Doris Jonas and Foster, Henry H., Jr., "Divorce in the 50 States: An Overview as of 1978," *Family Law Quarterly*, Spring 1979, *13*, 105–28.

Hall, Robert E. and Lazear, Edward P., "The Excess Sensitivity of Layoffs and Quits to Demand," *Journal of Labor Economics*, April 1984, *2*, 233–57.

Hart, Oliver D., "Optimal Labor Contracts Under Asymmetric Information: An Introduction," *Review of Economic Studies*, January 1983, *50*, 3–35.

Hashimoto, Masanori and Yu, Ben T., "Specific Capital, Employment Contracts and Wage Rigidity," *Bell Journal of Economics*, Autumn 1980, *11*, 536–49.

Johnson, William R. and Skinner, Jonathan, "Labor Supply and Marital Separation," *American Economic Review*, June 1986, *76*, 455–69.

Landes, Elisabeth M., "Economics of Alimony," *Journal of Legal Studies*, January 1978, *7*, 35–63.

Mincer, Jacob, "Family Migration Decisions," *Journal of Political Economy*, October 1978, *86*, 749–73.

Mnookin, Robert H. and Kornhauser, Lewis, "Bargaining in the Shadow of the Law: The Case of Divorce" *Yale Law Journal*, April 1979, *88*, 950–97.

Mortensen, Dale, "Specific Capital and Labor Turnover," *Bell Journal of Economics*, Autumn 1978, *9*, 572–86.

Patel, Jagdish K. and Read, Campbell B., *Handbook of the Normal Distribution*, New York: Marcel Dekker, 1982.

Peters, H. Elizabeth, "The Impact of Regulation of Marriage, Divorce and Property Settlements in a Private Contracting Framework," unpublished doctoral dissertation, University of Chicago, December

1983.

Riordan, Michael H., "Uncertainty, Asymmetric Information and Bilateral Contracts," *Review of Economic Studies*, January 1984, *51*, 83–93.

Schoen, Robert, Greenblatt, Harry N. and Mielke, Robert B. "California's Experience with Non-adversary Divorce," *Demography*, May 1975, *12*, 223–44.

Seal, K., "A Decade of No-fault Divorce: What It Has Meant Financially for Women in California," *Family Advocate*, Spring 1983, *1*, 10–15.

Sepler, Harvey J., "Measuring the Effects of No-Fault Divorce Laws Across Fifty States: Quantifying a Zeitgeist," *Family Law Quarterly*, Spring 1981, *15*, 65–102.

Sorensen, Annemette, "Flight from Unhappiness: Causes and Implications of the Recent Upturn in Divorce, The Case of Denmark," unpublished doctoral dissertation, University of Wisconsin-Madison, 1980.

Weitzman, Lenore J., *The Marriage Contract: Couples, Lovers, and the Law*, New York: Free Press, 1980.

Welch, Charles E. III and Price-Bonham, Sharon, "A Decade of No-fault Divorce Revisited: California, Georgia, and Washington," *Journal of Marriage and the Family*, May 1983, *45*, 411–18.

Wright, Gerald C., Jr. and Stetson, Dorothy M., "The Impact of No-Fault Divorce Law Reform on Divorce in the American States," *Journal of Marriage and the Family*, August 1978, *40*, 575–80.

U.S. Bureau of the Census, BLS, *Current Population Survey*, Match File, Washington, March/April 1979.

Labor Supply and Marital Separation

By WILLIAM R. JOHNSON AND JONATHAN SKINNER*

Panel data suggest that women who subsequently divorce increase their labor supply in the three years prior to separation. A simultaneous model of future divorce probability and current labor supply is estimated for married women. The results support the hypothesis that divorce probabilities increase labor supply. Thus, the recent rise in the frequency of divorce may account for one-third of the unexplained increase in women's postwar labor force participation.

One of the fundamental recent changes in the structure of the American family has been the rise in the frequency of divorce. The proportion of the population that had divorced and not remarried increased from 2.3 percent in 1960 to 6.7 percent in 1981. Marriages begun in 1970 were 60 percent more likely to end in divorce than those begun in 1950 (Andrew Cherlin, 1981). Despite these pronounced trends, the effect of marital separation on labor force participation and work hours has received little attention.[1]

Families in the Michigan Panel Study of Income Dynamics (*PSID*) that separated during one of the twelve survey years are usually the first to be excluded from empirical studies of labor supply. We use this neglected sample of families to investigate the effect of actual and possible marital separation on the labor supply of men and women. Panel data allow us to control for unobserved heterogeneity by measuring the labor supply of the same person in the years preceding and following the marital separation.[2] A comparison of average hours worked before and after the couples split indicates that actual separation marginally reduced work effort of men, but led to a substantial rise in female work hours. The average labor supply of women rose from 1024 hours one year before the split to 1551 hours four years following the separation.

Evidence from the *PSID* suggests that much of the total increase in female hours of work associated with the divorce occurred *before* the separation; work hours rose from an annual average of 744 six years preceding the split to 1024 the year before the separation. This shift in labor supply while the couple was still married is subject to two interpretations. First, women may respond to a higher probability of divorce by working to gain job experience; thus as the separation becomes more likely in the years preceding the split, hours of work rise. Alternatively, an increase in labor supply could raise the probability of divorce, so that women who had recently worked more hours would be overrepresented in the sample of divorced families.[3] We attempt to isolate

*Department of Economics, University of Virginia, Charlottesville, VA 22901. We thank Orley Ashenfelter, Tom Kniesner, H. Gregg Lewis, Glenn McDonald, Marjorie McElroy, Whitney Newey, Elizabeth Peters, Chris Robinson, Paul Schultz, Robert Shakotko, Steve Stern, John Strauss, several anonymous referees, and workshop participants at Duke, McMaster, Virginia, Western Ontario, and Yale for comments on earlier versions. Thanks also to Peggy Claytor for typing many drafts.

[1]Two exceptions are William Greene and Aline Questor (1982) and Robert Michael (1985). Michael uses aggregate time-series data to find that female labor force participation is significantly affected by lagged divorce rates, or that divorce "Granger causes" labor supply.

[2]While the *PSID* measures separation rather than divorce, we will use the two terms interchangeably.

[3]An alternative explanation would be that an unobservable effect (say, "political consciousness") led to both increased work hours and higher divorce rates. As long as this unobservable effect is not correlated with the exogenous variables, the validity of our estimates will not be affected.

these effects in a theoretical and empirical model in which divorce probabilities and labor supply are determined simultaneously. Our results show that married women without previous labor force experience increase their labor supply in response to higher divorce probabilities, and that working has a positive but insignificant effect on divorce.

Past studies of the labor supply of married women have left unexplained substantial variation in participation rates and in hours of work (see Thomas Mroz, 1985). The finding that labor supply is affected strongly both by marital separation and by the probability of future separation suggests that previous studies may have neglected an important factor in women's labor supply decisions. Most obviously, since divorce increases female labor supply, the rising fraction of women who are divorced acts to increase the overall labor force participation rate of women. More subtly, and more importantly, the rising probability of divorce faced by married women increases their labor supply, whether they ultimately separate or not. As we show, these effects could account for up to one-third of the unexplained increase in labor force participation among women in the past several decades. In addition, differences across groups in labor supply may in part reflect differences in divorce probabilities. For example, blacks have a higher probability of divorce than whites, even when other relevant variables are accounted for. This racial difference in divorce risk might explain a portion of the racial difference in female labor supply.[4]

Section I provides a descriptive overview of the relation between marital separation and hours of work. Section II sketches a model of divorce and labor supply, which is estimated in Section III using cross-section data from the *PSID*.

I. Hours of Work and Participation Rates for Women Who Divorce

During the first twelve years of the *PSID*, more than 20 percent of the sample experienced a change in marital status. As a first step, we examine changes in the labor supply of those families that experienced a marital separation during the survey period. The requirements for inclusion in the sample of families that separated were that they 1) separated between 1969 and 1977 (thus allowing at least one year of data in the survey before and after the separation), 2) remained separated for at least a year following the split, 3) were not retired, and 4) were not split-offs from the original sample. That is, if a teenager left the original family to get married in 1971, but separated in 1975, he or she would not be included in the sample. After checking for coding error and duplication of records, the sample consisted of 329 families.

Figure 1 (and Table 1) records average hours of work for women according to the number of years preceding or following the separation. These averages, which include women who did not work, show a slow rise in annual hours from the seventh year before the split (635 hours) to the year before the split (1024 hours), and then a sharp rise to around 1500 hours in the second year after the split. Much of the total increase in labor supply came through increased labor force participation, which rose from an average of 68 percent in the years before the split to 88 percent after the split.

The survey years were characterized by general increases in labor supply of all women, and it is instructive to compare the secular trend with these shifts in labor supply. We therefore include a baseline comparison of work hours for women who did not split during the twelve-year survey. These averages were adjusted to reflect the calendar year composition of labor supply for the separated women.[5]

[4]This explanation was first suggested by Glen Cain (1966). T. Paul Schultz (1980, p. 65) shows that racial differences in women's reservation wages persist after controlling for other characteristics.

[5]Translating calendar year averages into "year relative to split" averages was accomplished by appropriate weighting of the baseline calendar year averages. For

TABLE 1—HOURS OF WORK AND PARTICIPATION RATES OF SEPARATED COUPLES

	Females						Males	
	Full Sample			Average Hours				
Year Relative to Split	Average Hours	Participation Rate	N	($N=72$)[a]	($N=116$)[b]	Baseline Average Hours	Average Hours	N
−7	635	.62	58			595	2303	47
−6	744	.75	72	744		615	2219	59
−5	814	.67	96	748		616	2277	73
−4	760	.66	126	746		628	2267	82
−3	835	.67	149	766		637	2277	100
−2	921	.71	180	896		645	2321	119
−1	1024	.76	202	905	1113	664	2177	127
0	1250	.85	202	1236	1277	689	2259	127
+1	1420	.88	202	1445	1424	717	2151	127
+2	1507	.87	169		1501	729	1971	93
+3	1555	.86	137		1511	745	1952	69
+4	1551	.89	116		1551	763	1938	61

[a]Average hours for the sample of 72 families with data in year −6.
[b]Average hours for the sample of 116 families with data in year +4.

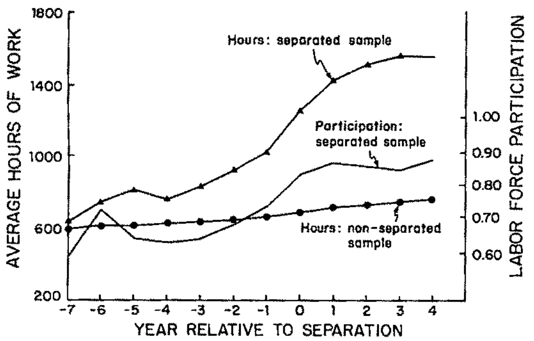


FIGURE 1. FEMALE HOURS OF WORK AND LABOR FORCE PARTICIPATION

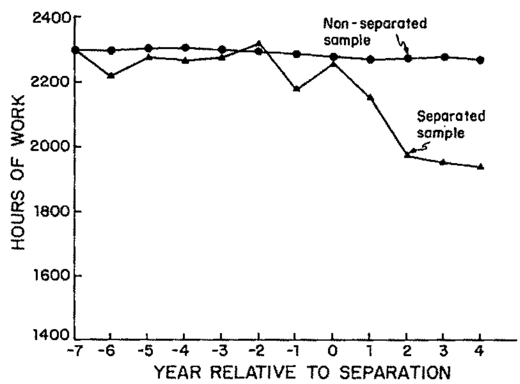


FIGURE 2. MALE HOURS OF WORK

One shortcoming of the data for separated couples is that the composition of the sample changes over the eleven years. Many of the divorced individuals remarried, and not all of the divorced spouses were followed by the Michigan Survey Center. To test whether this selection rule might have introduced some bias, we collected averages for groups that were married for at least six years previous to the split, and groups that were separated for at least four years after the split. The results in Table 1 suggest that this source of potential bias is not large. Correcting for the changing composition of the sample, however, reduces the variation in hours of work for earlier years (−6 to −3), and indicates that a large proportion of the rise in work effort for women occurred during the two years before the separation.

The *PSID* also indicates a large drop in family income for the recently divorced

example, if half the observations for 4 years before the split came from 1973 and half from 1974, then the appropriate comparison for the nonseparated sample is the simple average of labor supply in 1973 and in 1974.

woman. Income apart from her earnings dropped from \$9938 one year before the split to \$3032 the year of the separation. While consumption requirements may be less for single-headed households, it seems clear that the substantial loss in income provides an important incentive for increasing work hours. The magnitude of the income loss is roughly consistent with results in Lenore Weitzman (1985) and also supports the findings of Mary Jo Bane and David Ellwood (1986) that spells of poverty are often triggered by marital separation.

Hours of work for men are presented in Figure 2 and also in Table 1. Hours of work decline slightly in the year prior to the separation, rise during the year of separation, and then fall by about 10 percent after the separation. This fall in work hours is consistent with the theory of specialization in marriage; the loss of specialization causes the husband to spend more time working in household activities. Total work hours of the separated couple rises, owing both to the necessity for supporting two separate households, and to the loss of advantages from specialization.

The finding that women's work hours rise prior to the divorce suggests either that women anticipate working full time after the separation by gaining job experience early, or that the increases in the wife's working hours increases the risk of divorce. In the next section, these ideas are formalized in a simultaneous model of marital separation and married women's labor supply.

II. Labor Supply of Married Women with Endogenous Future Divorce

Consider the labor supply decision of a married woman with uncertain future marital status. Her current labor supply will have a positive effect on future wages because of learning by doing and on-the-job training. The present value of the investment in work effort, however, will differ according to the subsequent marital status of the wife. The economic theory of specialization in marriage suggests that were she to remain married with a wage less than her husband's, she would work fewer hours than if she were divorced. If she were divorced, both income effects and the loss of specialization would increase her labor supply, and hence raise the return to her accumulation of current job experience.[6] In other words, the expected value of marketable human capital should be greater the higher the probability of subsequent divorce. Hence we can write current labor supply, L, as a function of π, the probability of divorce, and a vector of conventional exogenous variables, X, explaining wife's labor supply or proxying for her market wage (for example, wife's education, husband's earnings, etc.):

$$L = g(\pi, X). \tag{1}$$

If the probability of divorce were exogenous, we could obtain estimates of (1) by using the actual occurrence of divorce in the future as a proxy for π. It is reasonable to suppose, however, that π is not exogenous and may, in fact, depend on the wife's current labor force behavior. According to the economic theory of marriage as espoused by Gary Becker (1981), the value of the marriage depends on the degree of specialization of the partners. Hence, holding other factors constant, the wife's decision to work will reduce the economic value of the marriage relative to divorce. Models of divorce and alimony developed by Becker, Elisabeth Landes, and Robert Michael (1977), Landes (1978), and Elizabeth Peters (1986) argue that divorce occurs when joint marital satisfaction falls below the joint level attainable by each partner separated.[7] If we suppose that new, unforeseen information becomes available in each period which may change the value of marriage to one or both partners, and this information is generated

[6] This abstracts from differences in marginal utility of income in the different states. See our earlier paper (1985).

[7] Transfers within marriage and the existence of alimony may allow efficient separation even if unilateral divorce is possible.

by a stochastic process which is identically and independently distributed across families, then divorce occurs when

$$h(L,Y)+\varepsilon<0, \tag{2}$$

where ε is the realized value of the new information, and h is the value of marriage, decreasing in L and also a function of other variables, Y. If ε is a standard normal deviate, then the probability of divorce, π, can be written

$$\pi=F(-h(L,Y)), \tag{3}$$

where F is the cumulative normal distribution function.

Since $L<0$ is not observed, equation (1) can be modeled as a Tobit process while divorce can be viewed as a probit process based on equation (3). To estimate the system consistently, we form a reduced-form probit for divorce based on all the exogenous variables of the system, $\{X \cup Y\}$. The predicted probability of divorce, $\hat{\pi}$, computed from the first-stage probit estimates is then used as a regressor in equation (1), which is estimated by maximum likelihood Tobit to account for the truncation of hours of work at zero. An identical procedure is also used to estimate a structural probit on labor force participation.

An alternative estimation method handles the truncation of hours at zero by first estimating an inverse Mills ratio based on a probit equation predicting labor force participation, and then using that ratio in an ordinary least squares (*OLS*) equation on hours of work. As John Cogan (1980) and others have shown, this procedure allows the index which determines hours of work to differ from the index determining labor force participation and is, therefore, more general than Tobit which assumes the two indices are identical.

III. Estimation of the Simultaneous Equation Model

To estimate the model, we used a cross-section sample of married families from the *PSID*. We chose 1972 for the sample year because it provided information on past work history in 1968–69 and information on future separations (between 1973 and 1978), and because a number of psychological questions which may be important in explaining divorce were asked by the survey in that year.[8] The sample combined two groups, all of whom had been married at least four years; 1599 couples who remained married during the entire panel period without any disruption in family structure, and 187 couples married in 1972 but who separated between 1973 and 1978. A description of the variables with their mean values is presented in Table 2. Many of the variables are standard to labor supply studies, such as age, education, and family income less the wife's earnings. We also include variables measuring psychological attitudes (an index of trust, an index of how easily one angers, whether one plans ahead), religious practices (frequency of religious attendance, religion), environment (whether close relatives live nearby, whether nonurban), and the tenure of the marriage.[9] Finally, a variable is included reflecting whether the state of residence had no-fault divorce laws in 1972.[10] This factor might affect both the probability of divorce and the financial support expected after divorce (Peters; Weitzman).

A complete model of household behavior would include current decisions about having children as another endogenous variable simultaneously determined with labor supply and divorce probabilities. However, *past* decisions about children and labor supply will also clearly affect current labor supply and

[8] These questions were asked only of one person in the household. Marital dissension might be closely tied, however, to *differences* in psychological attitudes between the spouses.

[9] Tenure of the marriage could only be measured for those in their first marriage. We know that those in their second marriage were married before 1968, so we assign tenure equal to 4 years to those who had remarried; this will tend to bias average tenure downward.

[10] The assignment of states was based on Peters' methodology applied to state divorce laws in 1972 as described by Doris Freed (1972).

TABLE 2—LIST OF VARIABLES WITH MEAN VALUES

Variable		Mean
1) Personal Characteristics		
Age (wife)	Age of wife	36.1
Age (husband)	Age of husband	38.9
Educ (wife)	Education (in years) of wife	11.7
Educ (husband)	Education (in years) of husband	11.5
Income	Family income minus wife's earnings ($\times 10^{-4}$)	1.153
Race	One if nonwhite (question asked in 1979)	.272
Disabled	Physical disability	.080
Infant	Whether youngest child is under age 3 in 1968	.275
Child	Whether youngest child is aged 3 to 8 in 1968	.346
2) Characteristics of the Marriage		
Tenure	Years of marriage (from 1968, updated to 1972)	14.06
Tenure^2	Tenure squared ($\times 10^{-2}$)	2.815
Second Mar	Current marriage at least second	.105
3) Environment		
Relatives	Whether close relatives live nearby	.472
Recent Move	Whether family recently moved	.176
Connected	Whether well connected with surrounding environment (index from 1 to 9; 9 is best)	6.33
Nonurban	Whether grew up outside urban area	.695
Divorce Laws	Whether lives in state with no-fault divorce in 1972	.242
4) Tastes		
Rel Attend	Frequency of religious attendance (1 = often)	.525
Baptist	Whether Baptist	.310
Catholic	Whether Catholic	.210
Jewish	Whether Jewish	.031
Plan Ahead	Whether plans ahead (index of 1 to 5; 5 is worst)	2.80
Horizon	Index of planning horizon (index of 1 to 8; 8 is most)	4.771
Anger Index	Whether gets angry often (index of 1 to 5; 1 is easily)	3.96
Trust Index	Whether trusts people (index of 1 to 5; 1 is most people)	2.35
Future	Whether thinks about future (index of 1 to 5; 1 is all the time)	3.19
Recreation	Type of recreational activity (1 is for productive, energetic activities, 0 for passive activities)	.765
5) Economic Variables		
Wage	Wife's wage rate (for those who work)	2.293
Labor Force Part	Wife's labor force participation in 1972 *Survey*	.572
Worked 68–69	Wife's labor force participation (0 if not worked in 1968–69, 0.5 if worked in 1968 or 1969, 1.0 if worked both years)	.470
Unemp Rate:	County unemployment rate (midpoint of brackets)	5.39

Note: Dichotomous variables, unless otherwise noted, are 1 if yes, 0 if no.

divorce probabilities. Whether such variables can be considered exogenous in our model is debatable. If one believes that the major source of randomness in the model is permanent unobservable differences in tastes or constraints across households, then past labor supply or fertility decisions are correlated with the current error term and these variables should not be included. On the other hand, if one believes that random influences on households are not strongly correlated over time, then past fertility and labor force events can be included as exogenous variables. We take an agnostic stance

on this issue by estimating two versions of the model, one with variables on children and labor force participation from 1968 to 1969 (four years before the 1972 survey) and one without such variables.

Table 3 compares selected variables for families who subsequently divorced, and for those who remained married through the 1970's. Average annual hours of work are higher for women who subsequently separate, owing mainly to their higher labor force participation rate. Conditional on working, annual hours of work between the two groups are similar; 1251 hours for wives in the separated sample, and 1241 for the sample that did not separate. There is little difference in work hours for the husbands. Couples that divorce tend to be about 3.5 years younger than those who do not, reflecting both cohort effects and the negative effect of marital tenure. Nevertheless, the separated sample represents a reasonably mature group of families, with an average wife's age at separation of about 36 years.

The two groups differ little by educational level or by residence in states with no-fault divorce laws. Families that subsequently divorced were more likely to have lower incomes net of wife's earnings, and more likely to be nonwhite, on their second marriage, and have lower marital tenure. Finally, while there is no difference between the two groups in the proportion having children under age 3 in 1968, families who separated were more likely to have their youngest child between age 3 and 8 in 1968 than those families that did not.

TABLE 3—MEANS AND STANDARD DEVIATIONS OF SELECTED VARIABLES IN 1972 BY SUBSEQUENT MARITAL STATUS[a]

Variable	Separated 1973–78 Mean	Not Separated 1973–78 Mean
Annual Hours (wife)	836	696
	(828)	(830)
Labor Force Part (wife)	0.668	0.561
	(0.472)	(0.496)
Worked 68–69 (wife)	0.543	0.461
	(0.441)	(0.443)
Annual Hours (husband)	2178	2242
	(736)	(653)
Age (wife)	32.86	36.49
	(7.33)	(9.99)
Age (husband)	35.89	39.26
	(8.65)	(10.26)
Education (wife)	11.86	11.69
	(2.45)	(2.74)
Education (husband)	11.45	11.48
	(3.18)	(3.64)
Income less Wife's Earnings	10,137	11,695
	(5,593)	(7,619)
Divorce Laws	0.246	0.241
	(0.432)	(0.428)
Tenure of Marriage	10.26	14.51
	(6.55)	(9.31)
Second Marriage	0.150	0.099
	(0.358)	(0.299)
Race	0.385	0.258
	(0.488)	(0.437)
Infant	0.369	0.263
	(0.484)	(0.441)
Child	0.342	0.346
	(0.476)	(0.476)
N of Cases	187	1599

[a]Standard deviations are shown in parentheses.

As explained in Section II, the first step of the estimation procedure is to estimate two reduced-form probit equations explaining future divorce and current labor force participation. The second-stage labor supply equation then uses the predicted divorce probabilities in the estimation of equation (1), while the second-stage divorce probit equation uses the predicted labor force participation in the estimation of (3).

The theoretical labor supply model includes variables such as age, education, race, family income less the wife's earnings, the county employment rate, and the state-specific divorce laws as explanatory variables. The psychological, religious, and marital tenure variables are thought not to affect labor supply decisions, thereby overidentifying the structural labor supply equation. We test these overidentifying restrictions later in this section. The county unemployment rate is assumed to affect only labor supply decisions, identifying the divorce equation.[11]

[11]Strictly speaking, the nonlinear transformation of the exogenous variables in the reduced-form labor force participation probit equation identifies the divorce equation, so even the country unemployment rate is not necessary for identification.

TABLE 4—REDUCED-FORM AND STRUCTURAL DIVORCE PROBIT EQUATIONS, 1972[a]

		Structural Model	
	Reduced-Form	Actual Labor Part.	Predicted Labor Part.
Variable	Coefficient	Coefficient	Coefficient
Age (wife)	−0.0001 (0.01)	−0.0007 (0.06)	−0.0006 (0.06)
Age (husband)	0.0059 (0.55)	0.0068 (0.63)	0.0076 (0.65)
Educ (wife)	0.0265 (1.25)	0.0226 (1.06)	0.0144 (0.33)
Educ (husband)	−0.0013 (0.07)	−0.0006 (0.04)	−0.0015 (0.09)
Income ($\times 10^{-4}$)	−0.0558 (0.65)	−0.0447 (0.52)	−0.0226 (0.17)
Race	0.2062 (1.77)	0.2068 (1.77)	0.1833 (1.33)
Relatives	−0.2183 (2.20)	−0.2113 (2.12)	−0.2024 (1.85)
Recent Move	0.1256 (1.13)	0.1248 (1.12)	0.1240 (1.11)
Disabled	0.4195 (2.95)	0.4137 (2.91)	0.3890 (2.19)
Horizon	0.1125 (2.72)	0.1109 (2.68)	0.1069 (2.43)
Connected	−0.0156 (0.50)	−0.0170 (0.55)	−0.0157 (0.51)
Tenure	0.0217 (0.73)	0.0243 (0.82)	0.0238 (0.78)
Tenure^2	−0.0018 (1.85)	−0.0018 (1.93)	−0.0018 (1.88)
Second Mar	−0.0620 (0.34)	−0.0499 (0.27)	−0.0324 (0.16)
Anger Index	−0.0026 (0.10)	−0.0016 (0.06)	0.0014 (0.05)
Trust Index	0.0362 (1.23)	0.0360 (1.22)	0.0349 (1.17)
Future	−0.0187 (0.45)	−0.0100 (0.41)	−0.0122 (0.50)
Recreation	−0.0724 (0.70)	−0.0721 (0.70)	−0.0937 (0.74)
Religious Attend	−0.4232 (4.33)	−0.4229 (4.32)	−0.4193 (4.22)
Baptist	0.2328 (2.15)	0.2195 (2.03)	0.2113 (1.71)
Catholic	0.0011 (0.01)	0.0044 (0.03)	0.0223 (0.16)
Jewish	−0.4642 (1.36)	−0.4771 (1.39)	−0.4812 (1.39)
Nonurban	−0.1642 (1.69)	−0.1631 (1.67)	−0.1651 (1.70)
Plan Ahead	0.0233 (0.94)	0.0232 (0.93)	0.0197 (0.71)
Divorce Laws	−0.0783 (0.76)	−0.0773 (0.75)	−0.0603 (0.48)
Infant	0.2096 (1.72)	0.2175 (1.78)	0.2409 (1.52)
Child	0.1218 (0.99)	0.1198 (0.97)	0.1177 (0.95)
Worked 68–69	0.2394 (2.32)	0.1803 (1.63)	0.0263 (0.04)
Labor Force Part		0.1334 (1.34)	0.4541 (0.31)
Unemp Rate	0.0081 (0.33)		
Constant	−2.1470 (4.18)	−2.1497 (4.29)	−2.1802 (3.95)
N of Cases	1786	1786	1786
Log-Likelihood	−531.03	−530.18	−531.03
R^2	0.076	0.076	0.076

[a] The *t*-statistics are shown in parentheses.

We first turn to the analysis of divorce probabilities. Table 4 presents the reduced-form and two structural divorce equations. The first structural equation uses actual 1972 labor force participation, while the second uses a fitted value for labor force participation estimated in an unreported reduced-form probit equation. There is little difference between the three equations with respect both to coefficients and to predictive properties.

The coefficients in the divorce probit equations can be converted to probability differences by a linear expansion of the normal distribution around the mean value of separation (a conversion factor of 0.2). Age or education of either the husband or the wife has little effect on divorce probabilities. The coefficient on race suggests that nonwhites have a higher divorce rate by approximately 4 percentage points, although the estimate is insignificant in the structural equation. The presence of nearby relatives reduces the probability of divorce by 4 percentage points, while being disabled increases the probability by nearly 8 percentage points.

The coefficients for tenure and tenure^2 suggest that the length of the marriage has at

first a positive, and then a negative effect on the chance of separation, with the maximum probability of divorce coming after ten years. Living in a nonurban area leads to a 3 percent lower probability of divorce but this coefficient is barely significant. Frequent religious attendance significantly reduces the chance of separating by approximately 8 percentage points. The particular religion is less important, although Baptists appear to have a higher and Jews a lower probability of separation. The presence of no-fault divorce laws is insignificant and reduces, rather than increases, the chance of separation.[12] Finally, both the fitted and the actual value of labor supply participation in the structural model have positive but insignificant effects on the probability of divorce.

Despite the low overall explanatory power of these three equations, the coefficients suggest substantial objective differences in expectations about divorce among families. For example, if we take two families otherwise alike, except that the first family does not attend church, has no relatives living nearby, lives in an urban area, and has a small child in the family, they will be 150 percent more at risk of separating in the next six years.

Turning to the analysis of women's labor supply, we present five Tobit equations on wife's hours of work in Table 5. The first equation uses a dummy for actual divorce with the limited set of variables and the second uses the fitted divorce probability with the same variables. They indicate that while age has little effect on labor supply, the wife's education has a strong positive effect on labor force participation and work hours. Nonwage income reduces labor supply by a significant degree (a $5000 increase in nonwage income reduces the probability of work by 6 percent, or reduces work hours by 200 hours). The unemployment rate has a negative effect, while being nonwhite has a positive effect on participation and hours of work. The no-fault divorce variable in the Tobit equations suggests a negative effect on labor supply in contrast to Peters' finding of a positive effect. Finally, both actual divorce and the probability of divorce have a positive though insignificant effect on work hours; the coefficient on divorce probability in column 2 translates to a change in labor force participation of 1.8 percentage points for each 10 percentage point rise in the probability of divorce.[13]

The third column in Table 5 presents the Tobit equations with the extended set of variables, including the presence of children and previous labor force participation. The divorce variable is also interacted with previous labor force participation to reflect differing returns to on-the-job experience. Since 82 percent of the women who had been working during 1968–69 continued to work during 1972, we hypothesize that divorce probabilities may have a differential effect on their work effort relative to the women who had never worked. Women without previous work experience may have the most to gain from accumulating marketable human capital.

Including work experience and children should provide a stronger test of our hypothesis since the probability of divorce would be likely to have affected past work and fertility behavior as well as current labor supply. As expected, both the third and fourth columns indicate that the presence of small children and past labor force experience strongly affect current labor supply. The coefficients on actual divorce (col. 3) are positive and are significant at the 0.01 level, while the interaction between past labor force participation and divorce is significantly negative and of a similar magnitude. These coefficients taken together imply that divorce affects labor supply for women with weak past work experience.

The estimated effect of marital separation on participation and hours of work is larger when the predicted (col. 4) rather than the actual (col. 3) divorce variable is used. The

[12] Peters also finds no significant effect of divorce laws on divorce probabilities.

[13] The divorce probabilities here were computed from an unreported reduced form on the limited set of variables.

TABLE 5—NORMALIZED COEFFICIENTS OF TOBIT LABOR SUPPLY EQUATIONS, 1972[a]

	Limited		Extended		Extended Model Fully Interacted	
	Actual (1)	Fitted (2)	Actual (3)	Fitted (4)	Variable (5)	Variable × Pr Divorce (6)
Age (wife)	0.0059	0.0071	0.0016	0.0043	0.0031	0.0531
	(0.88)	(1.05)	(0.24)	(0.62)	(0.26)	(0.66)
Age (husband)	−0.001	−0.0013	−0.0049	−0.0058	0.0006	−0.0361
	(0.16)	(0.20)	(0.74)	(0.88)	(0.05)	(0.50)
Educ (wife)	0.0765	0.0749	0.0614	0.0594	0.0328	0.1886
	(6.36)	(6.17)	(4.99)	(4.79)	(2.18)	(1.25)
Educ (husband)	0.0075	0.0076	0.0093	0.0093	0.0328	−0.2107
	(0.78)	(0.79)	(0.94)	(0.94)	(2.11)	(1.69)
Income ($\times 10^{-4}$)	−0.338	−0.3344	−0.2655	−0.2550	−0.1892	−0.8878
	(7.54)	(7.44)	(5.73)	(5.46)	(2.99)	(1.45)
Unemp Rate	−0.0437	−0.0439	−0.0218	−0.0237	−0.0044	−0.1875
	(3.05)	(3.05)	(1.49)	(1.62)	(0.19)	(1.23)
Race	0.2085	0.1885	0.2148	0.1854	0.3645	−1.3765
	(3.29)	(2.82)	(3.31)	(2.73)	(3.14)	(1.94)
Divorce Laws	−0.1426	−0.1427	−0.1397	−0.148	−0.1151	−0.2502
	(2.34)	(2.34)	(2.26)	(2.39)	(1.13)	(0.34)
Prob of (or actual) *Divorce*	0.1438	0.464	0.4255	2.0351	4.6146	
	(1.74)	(1.29)	(3.14)	(3.60)	(1.72)	
Infant			−0.2821	−0.3031	−0.3718	0.3013
			(4.02)	(4.22)	(2.94)	(0.35)
Child			−0.0659	−0.0663	0.0452	−1.103
			(1.06)	(1.05)	(0.46)	(1.40)
Work 68–69			1.2519	1.4241	1.4407	−2.565
			(18.95)	(14.65)	(14.65)	(3.59)
Work × Divorce			−0.5651	−2.4309		
			(2.99)	(3.45)		
Constant	−0.3192	−0.3699	−0.5188	−0.6878	−0.9925	
	(1.58)	(1.76)	(2.36)	(3.03)	(2.89)	
R^2	0.068	0.069	0.230	0.233	0.238	
Effect of .1 Divorce Increase on Participation (hours)	0.005 (18)	0.018 (58)	0.008[b] (22)	0.034[b] (97)	0.033[b] (92.2)	

[a]Absolute values of t-statistics are shown in parentheses.
[b]Divorce variables jointly significant at the 0.01 level.

coefficient on divorce rises to 2.03, and is highly significant.[14] This coefficient indicates that a .10 rise in the probability of divorce will lead to a .076 increase in labor force participation, or an increase of 123 annual hours of work for women who had no previous work experience. Once again, however, the interaction term between work experience and divorce suggests that the probability of divorce will have no effect on labor supply of women with previous experience.

The final two columns in Table 5 present the coefficients for a completely interacted regression equation. The interactions of divorce with previous labor force participation, and with race, are both negative and significant. These coefficients suggest that the

[14]Second-stage standard errors are biased downward because they are not adjusted for the fact that predicted probabilities are estimated with error by the first-stage procedure. The standard adjustment procedure, as described by Kevin Murphy and Robert Topel (1985) assumes that all of the error in the first-stage estimates is the error of the econometrician, not of the economic agent. However, in this case, the couple does not know with certainty whether they will divorce when they choose labor supply. Thus it is only to the extent that the couple's subjective probability explains more than our probit estimates that the standard errors are biased downward.

effect of divorce on labor supply is minimal or even negative for women either with past work experience or who are nonwhite. (We explore these behavioral differences further in Table 7 by separating the sample by race and by previous work experience.) The effect of an 0.10 increase in the probability of divorce for the combined sample is either a 3.3 percentage point increase in labor force participation, or a 92.2 hour yearly rise in labor supply.

The first six columns of Table 6 provide the coefficients for equations analogous to Table 5 but estimated using probit. The effect of actual divorce on labor force participation using the limited set of variables (col. 1) is significant, although the predicted effect of a 10 percentage increase in divorce is only 0.8 percent. Using predicted divorce in the limited model indicates a larger but insignificant effect of marital separation on labor supply. Turning next to the extended model with either actual (col. 3) or predicted (col. 4) divorce, the coefficient is positive and significant, with the coefficient on predicted divorce (1.542) roughly three times the coefficient for actual divorce (.443). One explanation for this difference might be that our estimated divorce probabilities measure true (subjective) divorce probabilities with less noise than actual subsequent divorce. The actual divorce variable would then introduce measurement error, causing a potential downward bias in its coefficient. For the entire sample of women, the estimated effect of a 10 percentage point change in divorce probabilities is 2.7 percentage points of labor force participation using the predicted values in column 4.

The fully interacted model of columns 5 and 6 shows significant negative interactive effects with race and previous work experience, and significant positive effects with the wife's education, with the interactive coefficients jointly significant at the 0.05 level. Finally, the last column in Table 6 presents a least squares regression for hours of work on the sample of women who worked in 1972. We adjust for selection bias by calculating an inverse Mills ratio from the reduced-form labor force participation equation, and use the log of her wage rate as an additional variable in the equation. The effect of divorce is substantial; a rise of 0.10 in divorce probabilities is predicted to increase work hours by 126.6 hours. However, the interactive term with past labor force experience indicates that potential divorce affects only women who have begun work in recent years. For those who had worked in the past, the effect is inconsequential.

Using the procedure described by Jerry Hausman (1978), we can test for the exogeneity of the divorce probability by including both the actual and the fitted value of divorce in the labor supply Tobit and probit equations. The test is whether predicted divorce and its interaction with past labor supply are jointly significant. While the coefficients are significant at the 0.05 level in the probit specification, the corresponding Tobit coefficients are not significant. Thus there is mixed evidence that divorce is endogenous, although the coefficients in Table 4 provide little support for the hypothesis that labor supply affects divorce.

A test of the overidentifying restrictions on the labor supply equation may be provided by comparing the performance of the structural equation with and without the excluded exogenous variables.[15] Under the null hypothesis of correct specification, the addition of the excluded variables should not appreciably affect the explanatory power of the equation. In the case of our labor supply probit equation, we find the null hypothesis rejected at the .10 level, but not at the .05 level using a standard likelihood-ratio test. In particular, two excluded variables are significant (recreation and whether plans ahead). Thus, the overidentifying restrictions do not seem to be seriously violated in this model, although the results may suggest a role for psychological variables in labor supply that economists typically ignore.

[15]Standard tests of overidentifying restrictions exist for linear models, but the nonlinear analogues are less straightforward (Hausman, 1983, p. 444). We are indebted to Whitney Newey for suggesting the test presented here.

TABLE 6—PROBIT AND *OLS* LABOR SUPPLY EQUATIONS, 1972

	Limited		Extended				
					Fully Interacted		
	Actual (1)	Fitted (2)	Actual (3)	Fitted (4)	Variable (5)	Variable ×Pr Div. (6)	Fitted *OLS* (7)
Age (wife)	0.0062	0.0071	0.0017	0.0042	−0.0028	0.0430	2.52
	(0.79)	(0.89)	(0.20)	(0.50)	(0.20)	(0.41)	(0.40)
Age (husband)	−0.0063	−0.0062	−0.0087	−0.0093	−0.0030	−0.0260	2.53
	(0.82)	(0.82)	(0.07)	(1.15)	(0.2)	(0.28)	(0.42)
Educ (wife)	0.0917	0.0903	0.0804	0.0784	0.0341	0.4900	6.50
	(6.38)	(6.24)	(5.17)	(5.01)	(1.43)	(2.46)	(0.76)
Educ (husband)	0.0023	0.0023	0.0044	0.0045	0.0454	−0.3943	3.04
	(0.20)	(0.21)	(0.36)	(0.37)	(2.30)	(2.33)	(0.20)
Income ($\times 10^{-4}$)	−0.3156	−0.3129	−0.2453	−0.2387	−0.1512	−1.4023	−154.99
	(6.17)	(6.10)	(4.40)	(4.25)	(2.01)	(1.76)	(3.11)
Unemp Rate	−0.0417	−0.0416	−0.0224	−0.0233	0.0070	−0.3069	−18.87
	(2.45)	(2.44)	(1.22)	(1.27)	(0.24)	(1.43)	(1.46)
Race	0.1429	0.1246	0.1322	0.1016	0.3482	−2.0375	161.82
	(1.89)	(1.55)	(1.61)	(1.18)	(2.38)	(2.13)	(2.71)
Divorce Laws	−0.1463	−0.145	−0.1634	−0.1704	−0.2282	0.6304	−51.81
	(2.03)	(2.01)	(2.12)	(2.21)	(1.82)	(0.65)	(0.90)
Prob of (or actual) *Divorce*	0.2081	0.4912	0.4431	1.5423	4.204		1266.20
	(2.03)	(1.13)	(2.82)	(2.36)	(1.19)		(2.28)
Infant			−0.253	−0.272	−0.3073	−0.0397	−165.54
			(2.89)	(3.04)	(1.99)	(0.03)	(2.28)
Child			−0.0081	−0.0109	0.116	−1.3747	−85.17
			(0.10)	(0.14)	(0.93)	(1.25)	(1.59)
Work 68–69			1.4041	1.5026	1.5099	−1.8167	429.35
			(17.11)	(12.37)	(12.22)	(1.97)	(2.26)
Work × *Divorce*			−0.622	−1.7887			−1385.00
			(2.58)	(2.01)			(2.15)
Log(*Wage*)							73.05
							(1.99)
Inv. Mills Ratio							50.72
							(0.28)
Constant	−0.3266	−0.3743	−0.6801	−0.8089	−1.1356		809.18
	(1.37)	(1.51)	(2.46)	(2.84)	(2.62)		(3.95)
Log-Likelihood or (R^2)	−1163.3	−1164.7	−983.99	−985.32	−974.22		(0.09)
Effect of .1 Divorce Divorce Increase on Part. (hours)	0.008[a]	0.019	0.006[b]	0.027[a]	0.033[b]		(37.7)[a]

[a] Divorce variables jointly significant at the .05 level.
[b] Divorce variables jointly significant at the .01 level. Absolute value of t-statistics in parentheses.

The significant interaction terms for race and past labor force experience suggest that the prospect of marital separation may have different effects on subgroups of women. To relax the assumption that other variables equally affect labor supply for each group, we estimate separate regressions for whites, nonwhites, those with past work experience, and those without past work experience, with the results presented in Table 7. The coefficient on predicted divorce for whites is 3.068, with a t-statistic of 3.79, while the coefficient for nonwhites is 1.814, with a t-statistic of 2.08. The interaction term with past labor force participation is strongly negative; for blacks, the effect of divorce on labor supply conditional on having worked in 1968–69 is −1.86. The predicted effect of 0.10 increase

TABLE 7—TOBIT REGRESSIONS BY RACE AND PREVIOUS WORK EXPERIENCE, 1972[a]

	Whites	Nonwhites	1968–69 Worked	1968–69 Did Not Work
Age (wife)	−0.0076	0.0134	0.0139	−0.0074
	(0.87)	(1.87)	(1.24)	(0.66)
Age (husband)	0.0075	−0.0198	0.0064	−0.0188
	(0.89)	(1.82)	(0.62)	(1.69)
Educ (wife)	0.0686	0.0506	0.0432	0.0823
	(4.36)	(2.45)	(2.33)	(3.46)
Educ (husband)	0.0119	0.0082	0.0008	0.0151
	(0.92)	(0.52)	(0.05)	(0.87)
Income ($\times 10^{-4}$)	−0.2307	−0.4463	−0.3003	−0.1849
	(4.54)	(3.32)	(3.95)	(2.45)
Unemp Rate	−0.0306	−0.0211	−0.0529	−0.0134
	(1.81)	(0.70)	(2.33)	(0.52)
Race			0.206	0.1694
			(2.02)	(1.39)
Divorce Laws	−1.087	−0.2349	−0.3409	−0.0987
	(1.51)	(1.89)	(3.59)	(0.89)
Probability of Divorce	3.068	1.8142	0.1692	1.169
	(3.79)	(2.08)	(0.32)	(1.68)
Infant	−0.4687	−0.0602	−0.0855	−0.6324
	(5.38)	(0.45)	(0.74)	(5.02)
Child	−0.0446	−0.1033	0.0174	−0.077
	(0.62)	(0.80)	(0.18)	(0.71)
Work 68–69	1.3346	1.8597		
	(11.71)	(8.63)		
Work × Divorce	−2.7007	−3.6809		
	(2.72)	(3.18)		
Constant	−0.9134	−0.1953	0.4973	−0.177
	(3.36)	(0.45)	(1.48)	(0.42)
R^2	0.234	0.257	0.074	0.135
Effect of .1 Divorce Increase on Participation (hours)	0.07[a]	−0.002[b]	0.004	0.043
	(75.6)	(−2.5)	(3.9)	(57.5)

[a]Absolute value of t-statistics are shown in parentheses.
[b]Divorce variables jointly significant at the 0.05 level.
[c]Divorce variables jointly significant at the 0.01 level.

in divorce probabilities on participation (hours) is 7 percent (75.2) for whites and −.002 (−2.5) for blacks.

The last two columns in Table 7 present the labor supply equations by past labor force status. The effect of divorce either on participation or on hours of work is small and insignificant for those who had worked and larger but still insignificant for those who had not.

Jacob Mincer and Solomon Polachek (1974; 1978) refer to a number of studies which find that human capital of more educated women depreciates more from disuse than the human capital of less educated women. If future benefits from current market participation are higher for more educated women, we would expect to see a positive interaction between education and the coefficient on divorce probabilities. While the fully interacted probit equation in Table 6 supports this hypothesis, a set of regressions run separately for high school and non-high-school graduates (regressions not reported) do not. For the high school graduates, the Tobit divorce coefficient is 2.24 with a t-statistic of 3.14, while for non-high-school graduates, the estimated effect is

2.46 with a t-statistic of 2.54. In both regressions, the large negative interaction term with past work experience implies that the divorce probability affects work decisions of women with little previous work experience.

The empirical results presented in this section support the hypothesis that divorce probabilities exert a substantial effect primarily on the labor supply of women with little past work experience. For those who had worked in 1968–69, the probability of marital separation had a marginal or even negative effect on participation and work hours. The effect of divorce on participation and work hours was also larger for whites relative to blacks. The regressions in Table 4 provide little support for a positive effect of labor force participation in 1972 on subsequent divorce during 1973–78. However, a specification test rejects the null hypothesis that divorce is exogenous for the probit, but not the Tobit specification.

IV. Conclusion

We have found a significant effect of divorce probabilities on the labor supply of married women. This finding lends support to the interpretation of the labor supply pattern described in Section I as one in which women, anticipating a higher probability of separation, work more. There is little support for significant effects of labor force participation on divorce probabilities.

One key assumption in these estimates is that the exogenous factors are not themselves affected by unobservables. For example, if unobservable tastes affect women's labor force participation, divorce, and, say, church attendance, then we might estimate a spurious effect of the predicted divorce rate on labor supply. Future research, with better data, might be able to clarify this issue. We have also left unexplained the extent to which the effect of actual divorce on female labor supply is an income effect of lost husband's earnings, or caused by the loss in household specialization. The answer to this question is important in assessing the effect of public policies on divorced women. Finally, a more complete model of marriage and divorce would allow for bargaining strategies in determining family labor supply and consumption (as in Marjorie McElroy and Mary Jean Horney, 1981), and a theory of why the marriages are formed in the first place.

We can use our results to address the puzzling question of why female labor supply has risen substantially in the postwar era, a rise that cannot be entirely explained by standard economic factors (William Bowen and T. Aldrich Finegan, 1969; David Shapiro and Lois Shaw, 1983; James Smith and Michael Ward, 1985). The effect on women's labor supply of higher divorce rates over time is twofold. The first, direct, effect is to increase the share of divorced women among all women, while the second is to increase the effect of higher divorce probabilities among married women. The results in this paper allow us to make rough calculations of the contributions of each of these effects to the overall rise in women's labor force participation.

Table 1 shows that the effect of actual divorce on the probability of participation is approximately 20 percentage points (from about .68 to about .88). Since the fraction of all women currently divorced rose from 2.3 to 6.7 percent from 1960 to 1980, the direct effect of the rise in divorce is to increase women's labor force participation by 0.8 percentage points. We can compute the indirect effect by using the results from the fully interactive Tobit or probit regressions. Since a marriage begun in 1970 has a 60 percent higher divorce probability than a marriage begun in 1950 (Cherlin), a rough guess would be that a typical married woman in 1980 faced a 60 percent higher chance of divorce than her counterpart in 1960. Assuming that the same growth rate applies to our measure of divorce risk,[16] the rise in divorce probabilities between 1960 and 1980 implies, using the estimates of the extended model, an increase in female labor force participation of roughly 1.8 percentage points. Altogether the rise in divorce rates from 1960 to 1980 may explain up to 2.6

[16] This is a conservative calculation since the annual divorce rate more than doubled between 1960 and 1980.

percentage points of the overall 15 percent rise in women's labor force participation in that period. Since roughly half of this overall rise has been explained by previous cross-sectional studies, the change implied by our finding may account for one-third of the remaining unexplained increase in labor supply.

REFERENCES

Bane, Mary Jo and Ellwood, David T., "Slipping Into and Out of Poverty: The Dynamics of Spells," *Journal of Human Resources*, Winter 1986, *21*, 1–23.

Becker, Gary, *A Treatise on the Family*, Cambridge: Harvard University Press, 1981.

______, Landes, Elisabeth and Michael, Robert, "An Economic Analysis of Marital Instability," *Journal of Political Economy*, November 1977, *85*, 1141–88.

Bowen, William and Finegan, T. Aldrich, *The Economics of Labor Force Participation*, Princeton: Princeton University Press, 1969.

Cain, Glen, *Married Women in the Labor Force*, Chicago: University of Chicago Press, 1966.

Cherlin, Andrew, *Marriage, Divorce, Remarriage*, Cambridge: Harvard University Press, 1981.

Cogan, John, "Labor Supply with Costs of Labor Market Entry," in James Smith, ed., *Female Labor Supply*, Princeton: Princeton University Press, 1980.

Freed, Doris, "Grounds for Divorce in the American Jurisdictions," *Family Law Quarterly*, Summer 1972, 2, 179–212.

Greene, William H. and Questor, Aline O., "Divorce Risk and Wives' Labor Supply Behavior," *Social Science Quarterly*, March, 1982, *63*, 16–27.

Hausman, Jerry, "Simultaneous Equation Models," in Z. Griliches and M. Intriligator, eds., *Handbook of Econometrics*, Vol. I, Amsterdam: North-Holland, 1983.

______, "Specification Tests in Econometrics," *Econometrica*, November 1978, *46*, 1251–72.

Johnson, William and Skinner, Jonathan, "Labor Supply and Marital Separation," University of Virginia, 1985.

Landes, Elisabeth, "Economics of Alimony," *Journal of Legal Studies*, January 1978, *7*, 35–63.

McElroy, Marjorie and Horney, Mary Jean, "Nash-Bargained Household Decisions: Toward a Generalization of the Theory of Demand," *International Economic Review*, June 1981, *22*, 333–49.

Michael, Robert, "Consequences of the Rise in Female Labor Force Participation Rates: Questions and Probes," *Journal of Labor Economics*, January 1985, *3*, S117–46.

Mincer, Jacob and Polachek, Solomon, "Family Investment in Human Capital: Earnings of Women," *Journal of Political Economy*, March/April 1974, *82*, S76–108.

______ and ______, "Women's Earnings Reexamined," *Journal of Human Resources*, Winter 1978, *13*, 118–34.

Mroz, Thomas, "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistic Assumptions," University of Chicago, 1985.

Murphy, Kevin and Topel, Robert, "Estimation and Inference in Two-Step Econometric Models," *Journal of Business and Economic Statistics*, October 1985, *3*, 370–79.

Peters, H. Elizabeth, "Marriage and Divorce: Informational Constraints and Private Contracting," *American Economic Review*, June 1986, *76*, 437–54.

Shapiro, David and Shaw, Lois, "Growth in the Labor Force Attachment of Married Women: Accounting for Changes in the 1970's," *Southern Economic Journal*, April 1983, *50*, 461–73.

Shultz, T. Paul, "Estimating Labor Supply Functions for Married Women," in James B. Smith, ed., *Female Labor Supply*, Princeton: Princeton University Press, 1980, 25–89.

Smith, James and Ward, Michael, "Time-Series Growth in the Female Labor Force," *Journal of Labor Economics*, January 1985, *3*, S59–90.

Weitzman, Lenore J., *The Divorce Revolution*, New York: Free Press, 1985.

Evaluating the Effects of Optimally Distributed Public Programs: Child Health and Family Planning Interventions

By MARK R. ROSENZWEIG AND KENNETH I. WOLPIN*

Considerable public resources in both high- and low-income countries are devoted to the subsidization of fertility control and health investments. The effects of these programs are thus of some concern, and social scientists have devoted attention to the evaluation of these programs. Most evaluation studies (see, for example, Albert Hermalin, 1972; Mohammad Khan and Ismael Sirageldin, 1979; Rosenzweig and T. Paul Schultz, 1982; and our 1982 article) have essentially compared the variation in measures of the intensity of program effort across localities with the corresponding interarea variation in fertility and health. Little or no attention has been paid to the causes of the cross-area variability in the levels of such programs. Yet, if the allocation of public health and family planning services or subventions across localities is systematically related to factors determining fertility and health outcomes that are known to subsidy providers but unobserved by researchers, such cross-sectional estimates will produce misleading conclusions about program effectiveness.

To assess the biases in cross-area estimates of program effects and to formulate appropriate estimation strategies for such program evaluations requires an understanding of the rules by which subsidies are allocated and financed, and by which agents adjust their behavior to such interventions. While there have been recent econometric advances in the evaluation of training and income supplement programs in the presence of the self-selection of heterogeneous agents into or out of the programs (for example, Jerry Hausman, 1982; Nicholas Kiefer, 1979), there has been less interest in questions concerned with the existence of actual public programs and their distribution across heterogeneous localities. Moreover, existing models of private agent behavior incorporating endogenous public resource allocations do not provide much guidance for predicting how publicly financed human capital subsidies and, in particular, "family planning" subsidies are distributed among differentiated recipients. Altruism theories of public transfers (Harold Hochman and James Rodgers, 1969; Russell Roberts, 1984) would appear to suggest that the least-endowed receive the highest transfers, but such models provide no rationale for the use of subsidies to particular goods ("merit" goods) such as contraceptives. Pressure group theory (Gary Becker, 1983) suggests that groups that are (*i*) relatively small in number, (*ii*) have and can command resources for lobbying, and (*iii*) derive the greatest benefits from public transfers or interventions will receive the highest transfers. This model would appear to imply that the rich—small in number and with greater resources—rather than the poor would receive the highest fertility control subsidies, since, assuming that the poor have the largest families and face the same prices, they avert less births than the rich and thus benefit least from subsidies to fertility control.

While the well-documented existence of externalities from health associated with contagion and infection provides a Pigovian (and pressure group) rationale for the subvention of health investments among the low-health poor, the empirical and theoretical rationale for fertility control subsidization based on the existence of direct population exter-

*Department of Economics, University of Minnesota, Minneapolis, MN 55455, and Center for Human Resource Research, Ohio State University, Columbus, OH 43202, respectively. The research and data collection efforts underlying this study were funded in part by the U.S. Agency for International Development. Robert Evenson initiated the surveys of the Laguna households used in this research and collaborated with us in the 1979 resurvey. We are grateful for the comments of two referees.

nalities is less clear (Zvi Eckstein and Wolpin, 1985). Moreover, since a birth from any source contributes equally to population growth, the existence of the most obvious population externality, congestion, does not obviously provide a basis for selective subsidization of households by income or human capital endowments.[1] A behavioral model of public health and family planning subsidies is needed.

Empirical applications or tests of public allocation models have been scarce and have been principally concentrated in the area of agricultural policy (Joel Guttman, 1978; Wallace Huffman and J. A. Miranowski, 1981; Huffman and Mark McNulty, forthcoming). Such studies have used the pressure or interest group framework, in which the size and resources of the subsidized group (farmers) are viewed as the principal determinants of the group's subsidy. However, a fundamental weakness of this empirical literature is the implicit assumption that both a group's membership and its resources are not themselves functions of the magnitude of the subsidy—for example, if net transfers to the farming sector increase, more persons may become (or remain as) farmers. These studies do not resolve the essential identification problem in program assessment by merely switching left- and right-hand side variables. Nor do they examine the distribution of the transfer burden, or measure the magnitude or direction of the bias in evaluating the contributions of research and extension services to agriculture under the assumption that such programs are randomly allocated.

The possibility that subsidies to the specific activities of recipients alter the size of recipient population provides a possible rationale for not only the subvention of population control in the absence of direct population externalities or altruistic concern for family size per se, but also for the evident close correspondence between the distributions of public family planning and health program expenditures. In this paper we formulate and test an optimizing model determining the distribution of family planning and health subsidies across heterogeneous households, and assess the biases in cross-area estimates of the health effects of such subsidies due to public resource optimization. The model incorporates both health externalities and the endogenous response of the size of the recipient population to program subsidies.

In Section I, the model is set out and implications are derived for the optimality of family planning and health subsidies, and for their joint distribution. In Section II, longitudinal data describing child health and publicly provided family planning and health programs in twenty barrios in Laguna Province in the Philippines are used to estimate the effects of such programs on child health and the relationships between the distribution of the programs and preprogram health levels, that is, the governmental allocation rules. Consistent with the model, we find that (*i*) dates of family planning and health program initiation across barrios are positively correlated, (*ii*) family size and health are gross substitutes among households and, in some barrios, family planning programs but not health programs are present; (*iii*) both programs were initiated earliest in the low-health barrios, and (*iv*) as a consequence, the positive and significant child-health effects of both the family planning and health programs are completely obscured when no account is taken of the systematic associations between program placement and preprogram health endowments.

I. Modeling the Distribution of Health and Family Planning Subsidies

A. *Evaluating Subsidy Effects on Health Production Among Heterogeneous Households*

Consider a set of T low-income households each residing in a different health environment. Each household i chooses a level of health for its children H^i, its family

[1] Given technological advances in contraceptive technology, a rationale for the public dissemination of general contraceptive information may be warranted. However, this does not necessarily justify subsidization of contraceptive devices or of person-specific contraceptive services. Moreover, as low-fertility households gain most from the acquisition of contraceptive information, disproportionate information provision to such households would appear to be implied by interest group theory.

size N^i, and its consumption Z^i, solving the following problem:

$$(1) \qquad \max U^i = U(H^i, N^i, Z^i),$$

where health production is described by the function

$$(2) \quad H^i = h(X^i, N^i) + \mu^i,$$
$$h_x^i > 0,\ h_{xx}^i,\ h_{NN} < 0,$$

subject to the full-income constraint

$$(3) \quad F^i = p_N N^i + (p_c - s_c^i)(\nu^i - N^i) + (p_x - s_x^i) X^i N^i + p_z Z^i,$$

where X^i = per child health input; μ^i = exogenous, health environmental endowment; F^i = full income net of taxes, if any; ν^i = potential fertility in the absence of fertility control; p_K = price of good K, $K = N, X, Z$; p_c = cost of fertility control or averted births; and s_c^i and s_x^i are per unit subsidies to fertility control and health inputs, respectively, provided in each health environment by a central "government" or donor.[2]

The solution for each household's average per child health net of the environmental effect, in terms of the exogenous variables unique to it, is

$$(4) \qquad H^i = H(s_x^i, s_c^i, F^i, \mu^i).$$

Estimation of (4) to obtain the average effect of the subsidy s_j^i on child health when μ^i is unobserved yields the estimate:

$$(5) \quad \frac{dH^i}{ds_j^i} = h_x^i \frac{dX^i}{ds_j^i} + h_N^i \frac{dN^i}{ds_j^i} + \left(\frac{ds_j^i}{d\mu^i}\right)^{-1}, \qquad j = x, c.$$

The true effect of a change in the subsidy s_j on the health outcome is given by the first two terms in (5): the subsidy (price) effects on the health input provided to each child and on family size (the recipient population) weighted by their respective marginal health effects from (2). The third term in (5) is the bias which arises when the μ^i are unobserved by the researcher and vary with the subsidy rates. Only if the subsidies are distributed independently of the μ^i, or, more generally, of any of the parameters unique to each area which influence health investments, will the association between the subsidies and health net of observed variables provide unbiased estimates of subsidy effects.

The sign of the bias in (5) will obviously depend on the allocation rules used by the agents who distribute the subsidies. If such agents follow a compensatory rule, for example, providing higher subsidies to less-endowed areas, then the subsidy effects obtained from (4), estimated, say, by least squares, will understate the true consequences of increasing the subsidy for any randomly chosen household; if such subsidies go to the better endowed, their effect will be overestimated. Consideration of the possibly systematic association between subsidies provided to agents and the environmental or other characteristics of the agents in the estimation of subsidy effects clearly yields better (policy-relevant) estimates of those effects. Moreover, such estimates also permit the testing of models of governmental resource allocations, which should provide the rules by which public resources are distributed among heterogeneous agents or localities, as well as a rationale for the particular set of instruments used to effect resource transfers.

B. *The Optimality of Family Planning and Health Subsidies and their Distribution*

To discern the distributional patterns of subsidies to fertility control and health among households behaving as described above in the absence of any arbitrarily assumed direct population externality, consider a wealthy household having the same objective function as in (1), but facing a health

[2]We assume that households do not move across localities. The consequences of the selective migration of agents in response to changes in local programs are considered in our paper (1984b). There it is demonstrated that the consequences for program evaluation from selective migration depend on the source of heterogeneity—preferences and/or endowments—and on cross-price effects, as here. See fn. 8.

externality. In particular, let the technology of health production for the well-off household be

$$(6)\qquad H = h(X, N, H^*) \qquad h_X, h_{H^*} > 0,$$

where H^* is the mean health of the children in the T low-income households, that is,

$$(7)\qquad H^* = \sum_{i=1}^{T} \alpha^i H^i,$$

where $\alpha^i = N^i/(\sum_{i=1}^{T} N^i)$ and $\max(H^i) < H$. Thus, while the health of the well-off children depends on the mean health of the children of the poor, there is no direct fertility externality. As in altruism models, the externality is asymmetric—poor households do not consider or are not affected by the consumption set of the "donor" household.[3]

Assume that the wealthy household can observe all the health endowments, but cannot tax fertility directly ($s_c^i \geq 0$).[4] If each household's fertility control and health expenditure can be differentially subsidized, the budget constraint for the wealthy household is

$$(8)\quad G = P_N N + p_c(\nu - N) + p_x XN + p_z Z + \sum_i s_c^i(1+\theta_c^i)(\nu^i - N^i) + \sum_i s_x^i(1+\theta_x^i)X^i N^i$$

where G = full income of the high-income household and θ_c^i and θ_x^i are the losses in subsidy transfers to the ith household associated with transaction costs (waste, graft). In this setup, the transfer scheme is politically feasible, since the majority of households (the poor) and possibly all households are potentially better off. The questions are: (*i*) Under what conditions will the wealthy household subsidize health *and/or* fertility control (when there is no population externality)? (*ii*) How will the subsidies, if warranted, be distributed among the poor households?

Maximization of the wealthy household utility function subject to (6), (8), and the price-taking behavior of the poor households, as described by model (1) through (3), yields equilibrium conditions for the distribution of subsidies across households (see the Appendix). These indicate that there are four sources of potential benefits to the donor households from fertility control subvention when a health externality is present. The first arises from the existence of cross-price effects: raising the fertility control subsidy increases health if fertility and child health are gross substitutes. A second potential source of gain exists if there is a direct, negative biological effect of family size on child health, through (2), in poor households.[5] A third benefit, which does not depend on either the particular preferences of households or the health technology, is the "eradication" effect of fertility control—decreases in the size of families with below-average child health ($H^i < H^*$) increase the mean health of the poor households; family planning subsidies provided to the lowest-health households thus increase the health of the wealthy households via the health external-

[3] The implications of the model are not sensitive to the assumption that the health of children in poor families is affected by the mean health levels of all poor households. Note also that the model yields essentially equivalent results if the wealthy household is characterized by an altruistic concern for the (mean) health of children from poor households. We prefer to specify the technological health externality, since there is direct evidence of its existence and importance.

[4] If only health outcomes (not endowments) are observed, potential recipients of subsidies may engage in "gaming" behavior, withholding resources allocated to health in order to increase the transfers to them. We believe, however, that the discrepancy between information held by researchers and subsidy providers with respect to the spatial variability in exogenous health conditions is the more important information asymmetry.

[5] The evidence on the biological effect of family size or birth order on child health suggests that such a linkage provides little justification for subsidization of fertility control on health grounds. In Rosenzweig and Schultz (1983), birthweight is found to significantly *increase* with increasing birth order; in our paper (1984a), little or no relationship is found between birth order and birthweight, although longer (prior) birth intervals increase birthweight. Both of these studies take into account in estimation the existence of heterogeneity in health endowments.

ity. The fourth source of gain from the fertility control subsidy also is independent of preferences or technology. It arises whenever there is also a positive health subsidy, because of the interaction between family size and per child health expenditures in the "governmental" budget constraint (8)—an increase in the family planning subsidy to household i which lowers family size in i by one child saves the subsidizing agent the amount $s_x^i X^i$. The cost of increasing the family planning subsidy depends positively on the number of births averted by the ith household and on the magnitude of transaction costs, and varies inversely with the magnitude of the own-fertility control elasticity.

The gain to the donor from subsidizing the health expenditures of household i depends positively on the magnitudes of the health externality and of the ith household's own-price elasticity of the health input X, but does not depend directly on the size of the fertility control subsidy. However, the cost of increasing the health subsidy to household i depends on the number of averted births in i (on i's family size) and thus indirectly on the household's ability to control family size.

These results imply that (i) when the health externality is sufficiently large to justify a health subsidy, the existence and magnitude of the family planning subsidy will be positively related to the size of the health subsidy —when both are used, family planning and health subsidies will be positively correlated, (ii) when cross-price effects and direct family size-health interactions are absent, family planning subsidies will be optimal only if health subsidies are optimal; and (iii) family planning subsidies will therefore be optimal in the absence of health subsidies only if health and family size are gross substitutes (or increases in family size directly lower health).

It can be shown that in the case in which fertility control but not health subsidies are present, sufficient (but not necessary) conditions for larger family planning subsidies to be provided to *low*-endowment households (compensatory subsidization) are that (i) fertility and health are gross substitutes ($dX/dp_c < 0$, $dN/dp_x > 0$), and (ii) $dX^i/d\mu^i < 0$; that is, more-endowed households invest less in health. In that case, the returns to further health investments will be smaller in high- than in low-μ households and high-μ households will have at least as many averted births (at least as high family planning subsidy costs) as low-μ households (see the Appendix). When strictly positive health and family planning subsidies are jointly optimal, the magnitudes of the subsidies will also generally depend on the differing health endowments of the recipient households. Moreover, the direction of the endowment-subsidy association is likely to be identical for both the family planning and health subsidy. However, unlike in the single-subsidy case, no simple sufficient condition regarding household-demand relationships determines the sign of the associations between the two subsidies and the health endowments. Even in the simple optimizing model adopted here, it is difficult to assess a priori the biases in the evaluation of family planning and health programs.

II. Empirical Application: Laguna Province

A. *The Data and the Distribution of Government Facilities*

We have shown that the effects of government interventions on per child health within a family are incorrectly estimated if the distribution of those interventions is influenced by the health predispositions of households, associated with endowments or tastes, that are unobserved by the researcher. In order to correctly assess the impact of government programs designed to influence health outcomes and to discover the government placement rules, it is thus necessary either to estimate or to measure preprogram heterogeneity in health outcomes. We will attempt to obtain consistent estimates of both the health effects of governmental family planning and health facilities, and of facility placement rules based on longitudinal data describing the distribution of such public programs and child health in twenty barrios (villages) in the lowland rice-producing areas of Laguna Province in the Philippines. Information from surveys of 240 randomly selected households residing in these barrios

TABLE 1—DISTRIBUTION OF PUBLIC FACILITIES IN TWENTY LAGUNA BARRIOS BY NUMBER OF YEARS INSTITUTED PRIOR TO 1979

Years in Barrio	Family Planning Clinic	Rural Health Clinic
0	8	7
0–4	4	3
5–9	5	0
10–14	2	4
15–19	0	2
20+	1	4
Total	20	20

on the age, height, and weight of every family member was collected in 1975 and 1979. Information was also obtained in the 1979 survey round on the dates of introduction of rural health clinics and family planning clinics financed by the national government for each of the barrios.

The distribution of the public facilities across barrios by time period is reported in Table 1. As can be seen, health and family planning facilities were relatively recently introduced, with barrio-specific dates of establishment differing markedly. Seven barrios had no public health clinic and eight barrios had no family planning facility by 1979, with seven of the thirteen existing health facilities and eleven of the twelve family planning facilities introduced less than fifteen years prior to 1979.

The joint distribution of the family planning and health clinics appears in conformity with the health externality model, as such facilities appear to be placed in a complementary pattern—the Spearmen rank correlation of establishment dates for the family planning and health clinics is .62. Moreover, of the seven barrios that had no health clinic, five also did not have a family planning clinic, and of the eight barrios without a family planning clinic, five also did not have a health clinic. Five barrios had neither facility as of 1979. The existence of two barrios without a health clinic but with a family planning clinic suggests, as noted, that if direct population externalities are ruled out, child health and family size should appear to be gross substitutes among the Laguna households. This is confirmed below.

B. *Estimation Framework*

To exploit the longitudinal data on health and the information on the dates of program initiation, we modify the above framework to accommodate the realities that government programs are initiated at different times and that observed child health in any period is a stock variable influenced by resources allocated in the current and prior periods. The impact of a program on the current health status of a particular child will thus depend upon the length of its previous exposure to the program. We will exploit the variability in program exposure across children to estimate the health effects of the health and family planning programs, and to estimate the barrio-specific health endowments. Variation in program exposure across children, however, occurs both because barrios differ in the timing of program introduction and because children within the barrio differ in their dates of birth. If child-health investments differ systematically by the birthdate of the child due to health-related factors about which the researcher is unaware, a spurious relationship between child health and program exposure is generated even if the timing of government programs across barrios is unrelated to family or barrio endowments.

Let t_s represent the year of the survey, t_p the year the program was instituted, and t_b the year of birth of a child. The program will have been in effect $t_s - t_p$ years and for children born prior to t_p (i.e., $t_b < t_p$), $t_s - t_p$ will be the number of years each such child will have been exposed. Yet, a child born one year prior to the program will likely be more strongly affected by the program than a child born five years prior to the program. We thus adopt as a measure of program exposure the fraction of a child's lifetime during which the child was exposed to the program. Let p_{il}^a be the program exposure of child i residing in barrio l who is of age a at the survey date, where $a = t_s - t_b$. Thus,

$$\begin{aligned} p_{il}^a &= 0 \text{ if the program does not exist in the barrio as of the survey date,} \\ &= ((t_s - t_p)/a) \text{ if } t_s \geq t_p > t_b \text{ in barrio } l, \\ &= 1 \text{ if } t_s > t_b \geq t_p \text{ in barrio } l. \end{aligned}$$

Consider the following child-health demand equation for a child i aged a in barrio l observed at t_s:

$$H^a_{ilt_s} = p^a_{il}\beta + u_i + \mu_l + \varepsilon_{ilt_s}, \tag{9}$$

where H is an age-standardized measure of health, u_i is a time-invariant, child-specific health endowment, the μ_l are location-specific health factors, and ε is a random error term. Least squares estimation of (9) when μ_l is unobserved leads to a biased estimate of β, the program exposure effect, if t_p, the date the program was introduced, is related to the area's endowments, as would be the case with nonrandom program placement.

Within-family or barrio estimators of β, which purge out, respectively, household and locational characteristics, are also biased, however, even if program placement is uncorrelated with child or family-specific endowments u if child-specific health endowments (within-family) or household endowments (within-barrio) influence the spacing of children. In differenced form, for a family with at least one child born prior to the program's introduction, the within-family estimator is derived from

$$H^{a'}_{jlt_s} - H^a_{ilt_s} = \left[\frac{(t_s - t_p)(t_{b'} - t_b)}{(t_s - t_{b'})(t_s - t_b)}\right]\beta + (u_j - u_i) + (\varepsilon_{jt_s} - \varepsilon_{it_s}), \tag{10}$$

where $a' = t_s - t_{b'} > a$. As can be seen, even if the dates of program introduction t_p are independent of the child-specific error u, or if child i's birth date t_b is related to his or her older sibling's health status u_j, the within-family estimate of the program exposure effect is also biased. In our paper (1984a) and Rosenzweig (1986), it is shown that birth spacing and other child-specific inputs are significantly correlated with *prior* siblings' and family-specific endowments, leading to biased estimates of child-specific resource allocations. Thus, as long as program placement is not responsive to purely random disturbances (or perturbations with little persistence), only within-child estimators will yield consistent estimates of the effect of program exposure, given systematic program placement and endowment-conditioned birth-spacing behavior.[6] Longitudinal information on child-health outcomes is required.[7]

C. *Program Assessment: Comparisons of Cross-Sectional and Panel Estimates*

To estimate the effects of the facilities on child health, we selected a sample of children (defined to be under age 18 as of 1979) for whom height and weight information exists in both years of the Laguna survey, yielding a working sample of 274 children in 85 households.[8] Table 2 provides descriptive

[6] Indeed, it is not necessarily true that the within-barrio regression performs better than the cross-section regression if the within-child regression is taken to be the correctly specified model.

[7] Similar biases would afflict least squares estimates of the effects of the two programs on fertility outcomes based on the cross-barrio variation in program exposure, as fertility patterns would not be independent of health endowments. Unlike the case of the health of individual children, however, within-family estimates are ruled out and use of longitudinal information based on birth histories would require that the timing and spacing behavior of households be better understood. Reductions in the cost of fertility control, while they would lower completed family size, do not have predictable effects (in the absence of better dynamic models of household behavior) on the intertemporal patterns of births—better control over fertility by couples may result in earlier or later childbearing. The higher rates of childbearing in the later life cycle stages by more educated women compared to less educated women in the United States (Rosenzweig and Daniel Seiver, 1982) suggests that examining differences in birth rates over arbitrary life cycle periods across women exposed differentially to the health and family programs is not a suitable procedure for discerning the programs' impact on completed family size. Use of the fixed-effect procedure to evaluate health effects, given reasonable age standardization, yields useful results, as a postponement of health investments in the face of permanent reductions in health costs and in fertility control costs would not appear to be an optimal strategy given the apparent cumulative and long-term effects of such investments.

[8] Two alternative sources of bias are selective mortality and selective household migration. With respect to the latter, of all the households interviewed in 1975 inclusive of those from which height and weight data

TABLE 2—SAMPLE STATISTICS

Variable	Mean	Standard Deviation
Natural Logarithm of Height Normalized by Philippines Age Standard		
1975	4.525	.0715
1979	4.543	.0566
Natural Logarithm of Weight Normalized by Philippines Age Standard		
1975	4.377	.147
1979	4.407	.147
Exposure to Public Health Unit, Fraction of Years		
1975	.456	.480
1979	.512	.451
Exposure to Family Planning Clinic, Fraction of Years		
1975	.162	.314
1979	.285	.333
Number of Years Rural Health in Barrio, 1979	10.0	10.3
Number of Years Family Planning Clinic in Barrio, 1979	6.45	13.67
Number of Barrios	20	
Number of Children	274	

statistics for the sample children at each of the two survey dates. Height and weight are standardized by age and sex according to a national schedule. The average child in this sample in each of the two survey years is somewhat over 90 percent as tall as the average Filipino child of the same age and sex, but only a little over 80 percent as heavy. However, the average child in the sample has evidently grown slightly in both dimensions relative to the standard between the two surveys.

Three separate specifications were estimated corresponding to alternative assumptions about unobservables in the determination of height and weight. In the first column of Table 3, ordinary least squares (*OLS*) regressions are reported using the 1979 cross section of 274 children. The second column repeats the cross-sectional regressions, but adds barrio dummy variables. The third column reports first-differenced regressions using the 1979 and 1975 (matched) samples. In panel A, the dependent variable is the log of standardized height; in panel B, the dependent variable is the log of standardized weight.

The differences in estimated program exposure effects across the specifications are striking for either health measure. In the height regressions, both the cross-section and barrio fixed-effect health and family planning clinic "effects" are generally negative with standard errors that are at least as large as the point estimates. The child fixed-effect (longitudinal) estimates, however, indicate that exposure to health and family planning clinics increases height, with the family planning effect statistically significant at the usual confidence levels and the health clinic effect marginally significant. The point estimates indicate that the height of a child for whom no health clinic existed would be 5 percent below that for a child always exposed to a clinic, while exposure to a family planning clinic increases height by 7 percent.

The weight regressions tell a very similar story: the cross-section and within-barrio associations between health clinic exposure and age-standardized weight are negative, while the child fixed-effect estimates, measured relatively precisely, indicate that exposure to either the health or family planning programs increases the weight-for-age of children. Here, however, the family planning effect is somewhat more robust to specification, although the effect of this program on child weight is underestimated by more than 100 percent when only the cross-sectional variation in program placement is utilized. The point estimates (last column) indicate that unit increases in health and

were collected, less than 10 percent were not present in the same barrio and thus were not surveyed in 1979. Interbarrio mobility in the Laguna province is not high. In a moderate-to-high mortality environment such as Laguna, child mortality may introduce some selectivity into estimates of the consequences of programs oriented toward children. To the extent that such programs facilitate the survival of children endowed with low health, cross-sectional comparisons of child health according to the presence or absence of such programs will lead to an underestimate of their beneficial effects. Using the fixed-effect procedure as we do, however, yields consistent estimates of the program effects as long as propensities to survive are impounded in the fixed-health endowment, are time-invariant attributes of children, and children do not differ importantly in the degree to which the programs augment their health.

TABLE 3—ESTIMATES OF THE EFFECTS OF EXPOSURE TO GOVERNMENTAL PROGRAMS ON THE STANDARDIZED HEIGHT AND WEIGHT OF CHILDREN

Variable	*OLS* Cross Section[a]	Fixed Effect Barrio	Fixed Effect Child
A. Log of Standardized Height			
Rural Health Unit Exposure	−.00473	−.0205	.0511
	(0.53)	(0.40)	(1.21)
Family Planning Exposure	−.0131	−.00913	.0710
	(1.12)	(0.27)	(3.32)
R^2	.0339	.1695	.0660[b]
F	1.88	2.12	9.61
$d.f.$	268	249	272
B. Log of Standardized Weight			
Rural Health Unit Exposure	−.0313	−.162	.0992
	(1.35)	(1.20)	(1.52)
Family Planning Exposure	.0263	.0803	.121
	(0.87)	(0.90)	(2.76)
R^2	.0337	.1401	.0500[b]
F	1.87	1.69	7.16

[a]Equation also includes the age and educational attainment of each parent.
[b]From first-differenced equation.

family planning clinic exposure increase age-standardized child weight by 9 and 12 percent, respectively.

D. *Program Placement Rules*

Whether child health status is measured by age-standardized height or weight, the estimates of the child-health effects of the family planning program purged of contamination by the endogeneity of program placement or birth spacing in Table 3 indicate that child health and family size are substitutes (and/or family size directly decreases health)—subsidies to fertility control evidently augment resource allocations to child-health investment among Laguna households. Thus, as we have shown, family planning clinics may substitute for health clinics in the presence of health externalities and/or may effectively complement health clinics even in the absence of other externalities, due to the direct relationship between the family size of recipients and the total costs of subsidizing health investments incurred by the donor or public agent.

In this section we seek to discern whether the dates of introduction of both the health and family planning clinics are systematically related to the average child-health endowment within a barrio, that is, we estimate the governmental program allocation rules. The child-specific effects that are estimated from (9) contain the elements u_i, μ_l, a constant, and the effects of all time-invariant determinants of height and weight (for example, mother's schooling), but net out the effects of the programs. However, since there are only two observations on each child, the estimated fixed effect measures the true preprogram child effect with error. Averaging child-specific effects within each barrio thus yields a measure (gross of time-invariant factors and random errors) of preprogram barrio level health presumably observed by the government, though only indirectly by us, and used by it to plan the timing of public program introduction. We have two such measures, corresponding to height and to weight.

Table 4 reports the estimates of the impact of the average barrio-level health endowment as measured by (the ln of) child height on the length of time in years that each of the programs—health clinics and family planning clinics—has been in existence in the barrio. There are thus twenty observations. Parental education levels were initially included as discussed above, but were jointly insignificant at conventional levels and so are

TABLE 4—ESTIMATES OF THE EFFECTS OF BARRIO CHILD-HEALTH CONDITIONS ON THE PLACEMENT OF GOVERNMENTAL PROGRAMS[a]

	Public Program					
	Rural Health Unit			Family Planning Clinic		
Endowment Measure	(1)	(2)	(3)	(1)	(2)	(3)
ln Height, Standardized[b]	−102 (0.92)	–	–	13.2 (0.09)	–	–
ln Height Effect	–	−145 (4.30)	–	–	−129 (2.30)	–
Predicted ln Height Effect	–	–	−199 (4.34)	–	–	−151 (1.91)
R^2	.0452	.507	.512	.0004	.228	.168

[a] Dependent variable = years since program was initiated.
[b] *OLS* coefficient.

excluded from the results actually presented. The first row uses actual mean height in the barrio and would only be correct if the programs themselves had no impact on height. The second row uses the barrio-average fixed effect computed from the child fixed-effect-height regression reported in column 3 of Table 3. The third row uses the predicted height fixed effect obtained from a first-stage regression in which the (ln) height fixed effect is regressed on the (ln) weight fixed effect, computed from the last column estimates of Table 3. The purpose of this latter procedure is to purge the estimate of the height fixed effect of measurement error under the assumption that height and weight are both measures of the same underlying health indicator.[9]

While the timing of program initiation for both programs appears unrelated to average child height in the barrios (row one), when the height effects of the programs are removed, as in the second and third rows, the estimates indicate that the health and family planning clinics were distributed systematically over time and, as expected, were allocated in a similar manner.[10] Moreover, the statistically significant, fixed-effect estimates imply a compensatory government allocation rule for both of the evidently complementary health programs. Barrios with lower pre-program health "endowments" evidently received both types of health-augmenting programs earlier.

The point estimates based on the predicted height measure indicate that where preprogram standardized height was 1 percent higher (about one-fourth of the standard deviation), the introduction of a health unit was retarded by about two years. The distribution of family planning clinics was almost as responsive to health endowment variation; their introduction was delayed by about one and one-half years for every percentage increase in standardized height. The compensatory program placement rule followed by the governmental authorities for the complementary health and family planning programs thus appears to have been responsible for the significant negative biases observed in the cross-sectional estimates of the effects of the two programs in Table 3.

III. Conclusion

In this paper we have specified a model of the distribution by a central authority of

[9]The results reported in Table 4 are qualitatively identical when the standardized weight effect is used to measure the community-level health endowment.

[10]An alternative explanation for the evident complementarity between family planning and health clinics is the existence of economies of scale in the delivery of related services. While this is plausible, both programs were integrated and/or conducted from the same building in less than half of those barrios where both types of clinics were present in 1979.

family planning and child health investment subsidies across heterogenous localities to assess the bias in the evaluations of such programs based on cross-sectional data implied by nonrandom program distribution. A basic feature of the model is the presence of health externalities, which is shown to be sufficient along with plausible features of household behavior to make selective subsidization of fertility control, either alone or in combination with health investment subsidies, Pareto efficient. Thus the unresolved issue of whether or not population growth per se impedes economic development, whether there are direct and negative population externalities, may be irrelevant to the issue of whether family planning programs are desirable instruments for promoting economic growth.

The model suggests that subsidization of fertility control is likely to be Pareto efficient in the presence of health or human capital externalities when (a) human capital and family size are gross substitutes, and/or (b) any per child human capital subsidies are provided to recipient households. In the first case, fertility control subsidies may substitute for direct subsidies to health investment and an equalizing distribution of the subsidies, the highest family planning subsidies to the lowest-health recipient households, is efficient. When both family planning and health subsidies are used, fertility control subsidies serve to minimize the subsidy burden for donors and will be highest where total subsidy expenditures per child are greatest, but in general the ordering of the distribution of the joint subsidies by the inherent healthiness of recipients cannot be predicted.

Longitudinal data describing the timing of program implementation in twenty randomly sampled barrios in the Laguna Province in the Philippines revealed a systematic pattern of health and family planning program placement in accord with the model: each program was initiated earliest in the low-health barrios, most of the barrios that had any program had both programs by the date of the last survey round, and when endogenous program placement was taken into account, exposure to either program appeared to significantly improve children's health status. Family size and child health thus appeared to be gross substitutes in the Laguna households, a sufficient condition for the presence of some barrios with a family planning, but not a health, clinic.

The compensatory pattern of program placement, when not taken into account, yielded estimates of the effects of the two programs on child health that would have led to false rejection of the hypothesis that either or both improved child health. These results thus imply that conclusions drawn from studies exploiting the cross-sectional variation in centrally allocated program intensities to evaluate programs and/or to identify structural relationships characterizing household behavior should be interpreted with care. Additional empirical and theoretical work integrating central and local program determination with household optimization would appear warranted.

Appendix

The equilibrium conditions for the optimal rates of subsidy for fertility control and health for the ith household, ignoring transaction costs, are

$$\text{(A1)}\quad \frac{s_c^i}{p_c} = -\alpha^i \frac{p_x N X \varepsilon_{HH^*}}{p_x N^i \varepsilon_{HX}} \left[\varepsilon_{HX}^i \frac{\eta_{Xp_c}^i}{\eta_{Np_c}^i} + \varepsilon_{HN}^i + \left(\frac{H^i - H^*}{H^*} \right) \right] - \left(n_{Np_c}^i \right)^{-1} \left(\frac{\nu^i - N^i}{N^i} \right) p_c + s_x^i \frac{X^i}{p_c} \left[1 + \frac{\eta_{Xp_c}^i}{\eta_{Np_c}^i} \right] \quad \text{or} \quad \frac{p_c}{s_c^i} = 0.$$

$$\text{(A2)}\quad \frac{s_x^i}{p_x} = \left[1 + \alpha^i \frac{NX}{N^i X^i} \frac{\varepsilon_{HH^*} \varepsilon_{HX}^i}{\varepsilon_{HX}} \times \left(\eta_{Xp_x}^i - \frac{\eta_{Np_x}^i \eta_{Xp_c}^i}{\eta_{Np_c}^i} \frac{\eta_{Np_x}^i}{\eta_{Np_c}^i} \frac{p_c}{p_x} \frac{(\nu^i - N^i)}{X^i N^i} \right] \psi^{-1} \quad \text{or} \quad s_x^i / p_x t = 0,$$

where

$$\psi=\left(\eta^i_{Np_x}+\eta^i_{xp_x}\right)\left[1-\frac{\eta^i_{Np_c}+\left(\eta^i_{Np_c}+\eta^i_{Xp_c}\right)}{\eta^i_{Np_c}\left(\eta^i_{Np_x}+\eta^i_{Xp_x}\right)}\right]$$

and $\varepsilon_{HH^*}=(\partial H/\partial H^*)(H^*/H)$ and $\varepsilon_{HX}=(\partial H/\partial X)(X/H)$, from (7); $\varepsilon^i_{HX}=(\partial H^i/\partial X^i)(X^i/H^i)$ and $\varepsilon^i_{HN}=(\partial H^i/\partial N^i)(N^i/H^i)$ from (2); and the η^i are the demand price elasticities characterizing the ith household. Note $\eta^i_{Np_c}>0$. The first term in brackets in (A1) is the health return from the fertility control subsity associated with the cross-price effect, the second term is that associated with the biological effect of family size on health and the third bracketed term corresponds to the eradication effect described in the text. The last term in (A1) is the budget savings effect resulting from the existence of the per capita health subsidy.

When the health subsidy is absent and $H^i-H^*=0$, total differentiation of the system of first-order conditions describing the wealthy household allocations, treating price effects as parameters, yields

$$\text{(A3)}\quad \frac{ds^i_c}{d\mu^i}=\left[\left(\frac{p_xN}{h_x}h^i_x\frac{(1-\alpha^i)}{\Sigma N^i}\left(\frac{dX}{dp_c}\right)-1\right)\times\left(\frac{dN^i}{d\mu^i}\right)-\alpha^i\frac{p_xN}{h_x}h^i_{xx}\left(\frac{dX^i}{dp_c}\right)\left(\frac{dX^i}{d\mu^i}\right)\right]\times\left(\frac{ds^i_c}{d\theta^i}\right)_c+\left[\frac{p_xN}{h_x}\alpha^i-s^i_c\left(\frac{dN^i}{d\mu^i}\right)\right]\frac{ds^i_c}{dG},$$

the effect of a change in the ith household's endowment on the fertility control subsidy it receives. The first bracketed term in (A3) corresponds to the (compensated) own-price or cost effect of the ith contraceptive subsidy; the second corresponds to the associated income effect on the wealthy household. When N^i and X^i are gross substitutes, $dN^i/d\mu^i<0$. If $dX^i/d\mu^i<0$ as well, then (A3) is negative and family planning subsidies are compensatory with respect to health endowments.

REFERENCES

Becker, Gary S., "A Theory of Competition Among Pressure Groups for Political Influence," *Quarterly Journal of Economics*, August 1983, *98*, 371–400.

Buchanan, James M. and Tullock, Gordon, *The Calculus of Consent*, Ann Arbor: University of Michigan Press, 1962.

Eckstein, Zvi and Wolpin, Kenneth I., "Endogenous Fertility and Endogenous Population Size," *Journal of Public Economics*, June 1985, *27*, 93–106.

Guttman, Joel, "Interest Groups and the Demand for Agricultural Research," *Journal of Political Economy*, June 1978, *86*, 467–84.

Hausman, Jerry A., "The Effects of Time in Economic Experiments," in Werner Hildebrand, ed., *Advances in Econometrics*, Cambridge: Cambridge University Press, 1982, 79–108.

Hermalin, Albert, "Regression Analysis of Areal Data," in C. Chandrasakaran and his *Measuring the Effects of Family Planning Programs on Fertility*, Belgium: Ordina Press, 1972.

Hochman, Harold M. and Rodgers, James A., "Pareto Optimal Redistribution," *American Economic Review*, September 1969, *59*, 542–47.

Huffman, Wallace E. and Miranowski, J. A., "An Economic Analysis of Expenditures on Agricultural Experiment Station Research," *American Journal of Agricultural Economics*, August 1981, *63*, 104–18.

______ and McNulty, Mark, "Endogenous Local Public Extension Policy," *American Journal of Agricultural Economics*, forthcoming.

Khan, Mohammad Ali and Sirageldin, Ismael, "Education, Income and Fertility in Pakistan," *Economic Development and Cultural Change*, October 1979, *27*, 519–48.

Kiefer, Nicholas M., "Population Heterogeneity and Inference from Panel Data on the Effects of Vocational Education," *Journal of Political Economy*, October 1979, *87*, S213–26.

Roberts, Russell D., "A Positive Model of Private Charity and Public Transfers," *Journal of Political Economy*, February 1984, *92*, 136–48.

Rosenzweig, Mark R., "Birth Spacing and Sibling Inequality: Asymmetric Information Within the Household," *International Economic Review*, February 1986, *27*, 55–76.

______ **and Schultz, T. Paul,** "Child Mortality and Fertility in Colombia: Individual and Community Effects," *Health Policy and Education*, February 1982, *2*, 305–48.

______ **and** ______, "Estimating a Household Production Function: Heterogeneity, the Demand for Health Inputs, and their Effects on Birthweight," *Journal of Political Economy*, October 1983, *91*, 733–46.

______ **and Seiver, Daniel,** "Education and Contraceptive Choice: A Conditional Demand Framework," *International Economic Review*, February 1982, *23*, 171–98.

______ **and Wolpin, Kenneth I.,** "Governmental Interventions and Household Behavior in a Developing Country," *Journal of Development Economics*, April 1982, *10*, 209–25.

______ **and** ______, (1984a) "Heterogeneity, Intrafamily Distribution and Child Health," Economic Development Center Discussion Paper No. 84–2, University of Minnesota, 1984.

______ **and** ______, (1984b) "Migration Selectivity and the Effects of Public Programs," Economic Development Center Discussion Paper No. 84–5, University of Minnesota, 1984.

Dividend Variability and Variance Bounds Tests for the Rationality of Stock Market Prices

By TERRY A. MARSH AND ROBERT C. MERTON*

Perhaps for as long as there has been a stock market, economists have debated whether or not stock prices rationally reflect the "intrinsic" or fundamental values of the underlying companies. At one extreme on this issue is the view expressed in well-known and colorful passages by Keynes that speculative markets are no more than casinos for transferring wealth between the lucky and unlucky. At the other is the Samuelson-Fama Efficient Market Hypothesis that stock prices fully reflect available information and are, therefore, the best estimates of intrinsic values. Robert Shiller has recently entered the debate with a series of empirical studies which claim to show that the volatility of the stock market is too large to be consistent with rationally determined stock prices. In this paper, we analyze the variance-bound methodology used by Shiller and conclude that this approach cannot be used to test the hypothesis of stock market rationality.

Resolution of the debate over stock market rationality is essentially an empirical matter. Theory may suggest the correct null hypothesis—in this case, that stock market prices are rational—but it cannot tell us whether or not real-world speculative prices as seen on Wall Street or LaSalle Street are indeed rational. As Paul Samuelson wrote in his seminal paper on efficient markets: "You never get something for nothing. From a nonempirical base of axioms, you never get empirical results. Deductive analysis cannot determine whether the empirical properties of the stochastic model I posit come close to resembling the empirical determinants of today's real-world markets" (1965, p. 42).

On this count, the majority of empirical studies report results that are consistent with stock market rationality.[1] There is, for example, considerable evidence that, on average, individual stock prices respond rationally to surprise announcements concerning firm fundamentals, such as dividend and earnings changes, and that prices do not respond to "noneconomic" events such as cosmetic changes in accounting techniques. Stock prices are, however, also known to be considerably more volatile than either dividends or accounting earnings. This fact, perhaps more than any other, has led many, both academic economists and practitioners, to the belief that prices must be moved by waves of "speculative" optimism and pessimism beyond what is reasonably justified by the fundamentals.[2]

*Sloan School of Management, MIT, Cambridge, MA 02139 and Hoover Institution, Stanford University; and Sloan School of Management, MIT, respectively. This paper is a substantial revision of a part of our 1983 paper which was presented in seminars at Yale and Harvard. We thank the participants for helpful comments. We also thank G. Gennotte, S. Myers, R. Ruback, and R. Shiller; and for his advice on the econometric issues, J. Hausman. We are especially grateful to F. Black, both for his initial suggestion to explore this topic, and for sharing with us his deep insights into the problem. We are pleased to acknowledge financial support from the First Atlanta Corporation for computer services. We dedicate the paper to the scientific contributions and the memory of John V. Lintner, Jr.

[1]To be sure, of the hundreds of tests of efficient markets, there have been a few which appear to reject market efficiency (see "Symposium on Some Anomalous Evidence on Capital Market Efficiency," *Journal of Financial Economics*, June-September 1978). For the most part, however, these studies are joint tests of both market efficiency and a particular equilibrium model of differential expected returns across stocks such as the Capital Asset Pricing Model and, therefore, rejection of the joint hypothesis may not imply a rejection of market efficiency. Even in their strongest interpretation, such studies have at most rejected market efficiency for select segments of the market. For further discussions, see Merton (1986).

[2]For example, in discussing the problems of Tobin's Q theory in explaining investment, Barry Bosworth

Until recently, the belief that stock prices exhibit irrationally high volatility had not been formally tested. In a series of papers (1981a, b, and 1982), Shiller uses seemingly powerful variance bounds tests to show that variations in aggregate stock market prices are much too large to be justified by the variation in subsequent dividend payments.[3] Under the assumption that the expected real return on the market remains essentially constant over time, he concludes that the excess variation in stock prices identified in his tests provides strong evidence against the Efficient Market Hypothesis. Even if the expected real return on the market does change over time, Shiller further concludes that the amount of variation in that rate necessary to "save" the Efficient Market Hypothesis is so large that the measured excess variation in stock prices cannot reasonably be attributed to this source.

We need hardly mention the significance of such a conclusion. If Shiller's rejection of market efficiency is sustained, then serious doubt is cast on the validity of this cornerstone of modern financial economic theory. Although often discussed in the context of profit opportunities for the agile and informed investor, the issue of stock market rationality has implications far beyond the narrow one of whether or not some investors can beat the market. As Keynes noted long ago (1936, p. 151), and as is evident from the modern Q theory of investment, changes in stock prices—whether rationally determined or not—can have a significant impact on real investment by firms.[4] To reject the Efficient Market Hypothesis for the whole stock market and at the level suggested by Shiller's analysis implies broadly that production decisions based on stock prices will lead to inefficient capital allocations. More generally, if the application of rational expectations theory to the virtually "ideal" conditions provided by the stock market fails, then what confidence can economists have in its application to other areas of economics where there is not a large central market with continuously quoted prices, where entry to its use is not free, and where shortsales are not feasible transactions?

The strength of Shiller's conclusions is derived from three elements: (i) the apparent robustness of the variance bound methodology; (ii) the length of the data sets used in the tests—one set has over 100 years of dividend and stock price data; and (iii) the large magnitude of the empirical violation of his upper bound for the volatility of rational stock prices. Shiller in essence relies upon elements (ii) and (iii) to argue that his rejection of the efficient market model cannot be explained away by "mere" sampling error alone.[5] Nevertheless, Marjorie Flavin (1983) and Allan Kleidon (1983a,b) have shown that such sampling error can have a nontrivial effect on the variance bound test statistics.

In this paper, we focus exclusively on element (i) and conclude that Shiller's variance bound methodology is wholly unreliable for the purpose of testing stock market rationality. Thus, even if his estimates contained no sampling error at all, his findings do not constitute a rejection of the efficient market model. To support our claim, we present an alternative variance bound test which has the feature that observed prices will, *of necessity*, be judged rational if they fail the

writes: "Nor does it seem reasonable to believe that the present value of expected corporate income actually fell in 1973–1974 by the magnitude implied by the stock-market decline of that period, when q declined by 50 percent. ...As long as management is concerned about long-run market value and believes that this value reflects 'fundamentals,' it would not scrap investment plans in response to the highly volatile short-run changes in stock prices" (1975, p. 286).

[3]Using the variance bound methodology, Stephen LeRoy and Richard Porter (1981) claim to show that stock prices are "too volatile" relative to accounting earnings. For a similar discussion of their analysis, see our 1984 paper.

[4]For a recent discussion of the "causal" effect of stock price changes on investment, see Stanley Fischer and Merton (1984).

[5]Shiller notes on this general point: "The lower bound of a 95 percent one-sided χ^2 confidence interval for the standard deviation of annual changes in real stock prices is over five times higher than the upper bound allowed by our measure of the observed variability of real dividends. The failure of the efficient markets model is thus so dramatic that it would seem impossible to attribute the failure to such things as data errors, price index problems, or changes in tax laws" (1981a, p. 434).

Shiller test. That is, if observed stock prices were to satisfy Shiller's variance bound test, then they would be deemed irrational by our test. It would seem, therefore, that for any set of stock market price data, the hypothesis of market rationality can be rejected by some variance bound test.

This seeming paradox arises from differences in assumptions about the underlying stochastic processes used to describe the evolution of dividends and rational stock prices. Affirmative empirical evidence in support of the class of aggregate dividend processes postulated in our variance bound test is presented in our forthcoming article. The specific model derived and tested in that paper significantly outperforms the univariate autoregressive model associated with the Shiller analysis.

The Shiller variance bound test and our alternative test share in common the null hypothesis that stock prices are rational, but differ as to the assumed stochastic process for dividends. Since Shiller's data sets strongly reject the joint hypothesis of his test and sustain our's, we conclude that his variance bound test results might better be interpreted as an impressive rejection of his model of the dividend process than as a rejection of stock market rationality.

I. On the Reliability of the Dividend Variance Bound Test of Stock Market Rationality

In his 1981a article, Shiller concludes that:

> measures of stock price volatility over the past century appear to be far too high—five to thirteen times too high—to be attributed to new information about future real dividends if uncertainty about future dividends is measured by the sample standard deviation of real dividends around their long-run exponential path. [p. 434]

In reaching this conclusion, he relies upon a variance bound test—hereafter called the "p^* test"—which establishes an upper bound on the variance of the level of detrended real stock prices in terms of the variance of a constructed "*ex post* rational" detrended and real price series.[6] In this section, we begin with a brief review of the development of his test and then present an alternative variance bound test which actually *reverses* the direction of the inequality established in the p^* test. That is, the *upper* bound on the variance of rationally determined stock prices in the Shiller test is shown to be the *lower* bound on that same variance in the alternative test.

The key assumptions underlying the p^* test can be summarized as follows:

(S1) Stock prices reflect investor beliefs which are rational expectations of future dividends.

(S2) The "real" (or inflation-adjusted) expected rate of return on the stock market, r, is constant over time.

(S3) Aggregate real dividends on the stock market, $\{D(t)\}$, can be described by a finite-variance stationary stochastic process with a deterministic exponential trend (or growth rate) which is denoted by g.

To develop the p^* test from these assumptions, Shiller defines an *ex post* rational detrended price per share in the market portfolio at time t:

$$p^*(t) \equiv \sum_{k=0}^{\infty} \eta^{k+1} d(t+k), \tag{1}$$

where $d(s) \equiv D(s)/(1+g)^{s+1}$ is the detrended dividend per share paid at the end of period s and $\eta \equiv (1+g)/(1+r)$. $p^*(t)$ is called an *ex post* (detrended) rational price because it is the present value of *actual* subsequent (to time t) detrended dividends. If as posited in (S1), actual stock prices, $\{P(t)\}$, are *ex ante* rational prices, then it follows from (1) that

$$p(t) = \varepsilon_t[p^*(t)], \tag{2}$$

[6] Shiller also develops a second variance bound test that establishes an upper bound on the variance of unanticipated changes in detrended real stock prices in terms of the variance of detrended real dividends. An analysis of this "innovations test" is presented later in this section.

for each t where $p(t) \equiv P(t)/(1+g)^t$ is the detrended real stock price per share of the market portfolio at the beginning of period t and ε_t is the expectation operator conditional on all information available to the market as of time t.

If, as Shiller (1981a, p. 422) points out, $p(t)$ is an *ex ante* rational price, then it is also an optimal forecast of $p^*(t)$. If $p(t)$ is such an optimal forecast, then the forecast error, $u(t) \equiv p^*(t) - p(t)$, should be uncorrelated with $p(t)$. It follows therefore that under this hypothesis, $\mathrm{Var}[p^*(t)] = \mathrm{Var}[p(t)] + \mathrm{Var}[u(t)] > \mathrm{Var}[p(t)]$. That is, in a set of repeated experiments where a forecast $p(t)$ and a sequence of subsequent dividends, $d(t+k)$, $k = 0, 1, \ldots$, are "drawn," it should turn out that the sample variance of $p^*(t)$ exceeds the sample variance of the forecast $p(t)$.

If (detrended) dividends follow a regular stationary process, then rationally determined (detrended) stock prices must also. Hence, from assumption (S3), it follows by the Ergodic Theorem that time-series ensembles of $\{p(t)\}$ and $\{p^*(t)\}$ can be used to test the "cross-sectional" proposition that $\mathrm{Var}[p^*(t)] > \mathrm{Var}[p(t)]$.[7]

To compute an estimate of $p^*(t)$ with a finite sample time period, it is, of course, necessary to truncate the summation in (1). If, as Shiller (1981a, p. 425) notes, the time-series sample is "long enough," then a reasonable estimate of the variance of $p^*(t)$ can be obtained from that truncated summation. At the point of truncation, Shiller assigns a "terminal" value, $p^*(T)$, which is the average of the detrended stock prices over the sample period. That is,

$$(3) \qquad p^*(T) = \left[\sum_{t=0}^{T-1} p(t)\right] \Big/ T,$$

where T is the number of years in the sample period.

Under the posited conditions (S1)–(S3), the null hypothesis of the p^* test for rational stock prices can be written as

$$(4) \qquad \mathrm{Var}[p^*] \geq \mathrm{Var}[p],$$

where from (1) and (3), the constructed $p^*(t)$ series used to test the hypothesis is given by

$$(5) \qquad p^*(t) = \sum_{k=0}^{T-t-1} \eta^{k+1} d(t+k) + \eta^{T-t} p^*(T), \quad t = 0, \ldots, T-1.$$

As summarized by Shiller in the paragraph cited at the outset of this section, the results reported in his Table 2 (1981a, p. 431) show that the variance bound in (4) is grossly violated by both his Standard and Poor's 1871–1979 data set and his modified Dow Industrial 1928–79 data set.

Although widely interpreted as a rejection of stock market rationality (S1),[8] these findings are more precisely a rejection of the joint hypothesis of (S1), (S2), and (S3). As noted in our introduction, Shiller (1981, pp. 430–33) argues that a relaxation of (S2) to permit a time-varying real discount rate would not produce sufficient additional variation in prices to "explain" the large magnitude of the violation of the derived variance bound. However, even if (S2) were known to be true, this violation of the bound is not a valid rejection of stock market rationality unless (S3) is also known to be true. Nevertheless, to some, (S3) may appear to encompass such a broad class of stochastic processes that any plausible real-world time-series of dividends can be well-approximated by some process within its domain.[9] If this were so, then, of course, the p^* test, viewed as a test of stock market rationality, would be robust. In fact, however, this test is very sensitive to

[7]That is, the time-series estimator $\Sigma_0^{T-1}[p(t) - \bar{p}]^2/T$ can be used to estimate $\mathrm{Var}[p(t)]$ and similarly for $p^*(t)$.

[8]As a recent example, see James Tobin (1984).

[9]Perhaps this belief explains why Shiller devotes 20 percent of his paper (1981a) to justifying the robustness of his findings with respect to assumption (S2) and virtually no space to justifying (S3).

the posited dividend process. We show this by deriving a variance bound test of rational stock prices that reverses the key inequality (4). While maintaining assumptions (S1) and (S2) of the p^* test, this alternative test replaces (S3) with the assumption of a different, but equally broad, class of dividend processes. As background for the selection of this alternative class, we turn now to discuss some of the issues surrounding dividend policy and the sense in which rational stock prices are a reflection of expected future dividends, this to be followed by the derivation of our test.

If the required expected real rate of return on the firm is constant, then its intrinsic value per share at time t, $V(t)$, is defined to be the present value of the expected future real cash flows of the firm that will be available for distribution to each of the shares currently outstanding. From the well-known accounting identity,[10] it follows that the firm's dividend policy must satisfy the constraint:

$$(6)\quad V(t)=\varepsilon_t\left[\sum_{k=0}^{\infty} D(t+k)/(1+r)^{k+1}\right].$$

Although management can influence the intrinsic value of its firm by its investment decisions, management has little, if any, control over the stochastic or unanticipated changes in $V(t)$. In sharp contrast, management has sole responsibility for, and control over, the dividends paid by the firm. There are, moreover, no important legal or accounting constraints on dividend policy. Hence, subject only to the constraint given in (6), managers have almost complete discretion and control over the choice of dividend policy.

This constraint on dividend choice is very much like the intertemporal budget constraint on rational consumption choice in the basic lifetime consumption decision problem for an individual. In this analogy, the intrinsic value of the firm, $V(t)$, corresponds to the capitalized permanent income or wealth of the individual, and the dividend policy of the firm corresponds to the consumption policy of the individual. Just as there are an uncountable number of rational consumption plans which satisfy the consumer's budget constraint for a given amount of wealth, so there are an uncountable number of distinct dividend policies that satisfy (6) for a given intrinsic value of the firm. Hence, like rational consumers in selecting their plans, rational managers have a great deal of latitude in their choice of dividend policy.[11]

If stock prices are rationally determined, then

$$(7)\qquad P(t)=V(t)\quad \text{for all } t.$$

Hence, the only reason for a change in rational stock price is a change in intrinsic value. Since a manager can choose any number of different dividend policies that are consistent with a particular intrinsic value of the firm, the statement that "rational stock prices reflect expected future dividends" needs careful interpretation. It follows from (6) and (7) that rational stock prices will satisfy

$$(8)\quad P(t)=\varepsilon_t\left[\sum_{k=0}^{\infty} D(t+k)/(1+r)^{k+1}\right].$$

Thus, rational stock prices reflects expected future dividends through (8) in the same sense that an individual's current wealth reflects his expected future consumption through the budget constraint. Pursuing the analogy further: if because of an exogenous event (for example, a change in preferences), a consumer changes his planned pattern of

[10] The cash flow accounting identity applies only to dividends paid *net* of any issues or purchases of its outstanding securities. "Gross" dividends are, of course, subject to no constraint. Hence, all references to "dividends" throughout the paper are to "net" dividends paid.

[11] The fact that individual firms pursue dividend policies which are vastly different from one another is empirical evidence consistent with this view.

consumption, then it surely *does not* follow from the budget constraint that this change in the expected future time path of his consumption will cause his current wealth to change. Just so, it does not follow from (8) that a change in dividend policy by managers will cause a change in the current rationally determined prices of their shares.[12] For a fixed discount rate, r, it does however follow from (8) that an unanticipated change in a rationally determined stock price must necessarily cause a change in expected future dividends, and this is so for the same feasibility reason that with a constant discount rate, an unanticipated change in a consumer's wealth must necessarily cause a change in his planned future consumption. *In short, (8) is a constraint on future dividends and not on current rational stock price.*

Since management's choice of dividend policy clearly affects the time-series variation in observed dividends, the development of the relation between the volatility of dividends and rational stock prices requires analysis of the linkage between the largely controllable dividend process and the largely uncontrollable process for intrinsic value.

Unlike the theory of consumer choice, there is no generally accepted theory of optimal dividend policy.[13] Empirical researchers have, therefore, relied on positive theories of dividend policy to specify their models. The prototype for these models is John Lintner's model (1956) based on stylized facts first established by him in a classic set of interviews of managers about their dividend policies. Briefly, these facts are: (L1) Managers believe that their firms should have some long-term target payout ratio; (L2) In setting dividends, they focus on the change in existing payouts and not on the level; (L3) A major unanticipated and nontransitory change in earnings would be an important reason to change dividends; (L4) Most managers try to avoid making changes in dividends which stand a good chance of having to be reversed within the near future. In summary, managers set the dividends that their firms pay to have a target payout ratio as a long-run objective, and they choose policies which smooth the time path of the changes in dividends required to meet that objective.

As most textbook discussions seem to agree, these target payout ratios are measured in terms of long-run sustainable ("permanent") earnings rather than current earnings per share. In the special case where the firm's cost of capital r is constant in real terms, real permanent earnings at time t, $E(t)$, are related to the firm's intrinsic value per share by $E(t) = rV(t)$.

With this as background, we now develop a model of the dividend process as an alternative to the p^* test's (S3) process. A class of dividend policies which captures the behavior described in the Lintner interviews is given by the rule:

$$(9)\quad \Delta D(t) = gD(t) + \sum_{k=0}^{N} \gamma_k [\Delta E(t-k) - gE(t-k)],$$

where Δ is the forward difference operator, $\Delta X(t) \equiv X(t+1) - X(t)$, and it is assumed that $\gamma_k \geq 0$ for all $k = 0,1,\ldots, N$. In words, managers set dividends to grow at rate g, but deviate from this long-run growth path in response to changes in permanent earnings that deviate from their long-run growth path. Describing the policies in terms of the

[12] By the accounting identity, net dividend policy (as described in fn. 10) cannot be changed without changing the firm's investment policy. However, changes in investment policy need not change the current intrinsic value of the firm. Managers can implement virtually any change in net dividends per share (without affecting the firm's intrinsic value) by the purchase or sale of financial assets held by the firm or by marginal changes in the amount of investment in any other "zero net present value" asset held by the firm (for example, inventories). Such transactions will change the composition of the firm's assets and the time pattern of its future cash flows, but not the present value of the future cash flows. Since these "trivial" changes in investment policy will not affect the intrinsic value of the firm, they will not affect the current level of rationally determined stock price. See our forthcoming article (Section 6.3) for further discussion of the difficulties of measuring net dividends.

[13] Indeed, the classic Miller-Modigliani (1961) theory of dividends holds that dividend policy is irrelevant, and hence, in this case, there is no optimal policy.

change in dividends rather than the levels, and having these changes depend on changes in permanent earnings, is motivated by Lintner's stylized facts (L2) and (L3). His behavioral fact (L4) is met in (9) by specifying the change in dividends as a moving average of current and past changes in permanent earnings over the previous N periods.

Equation (9) can be rewritten in terms of detrended real dividends and permanent earnings as

$$(10)\qquad \Delta d(t)=\sum_{k=0}^{N}\lambda_k\Delta e(t-k),$$

where $e(s)\equiv E(s)/(1+g)^s$ and $\lambda_k\equiv\gamma_k/(1+g)^{k+1}$. By integrating (10),[14] we can express the level of detrended dividends at time t in terms of current and past detrended permanent earnings as

$$(11)\qquad d(t)=\sum_{k=0}^{N}\lambda_k e(t-k).$$

By inspection of (11), the dividend policies in (9) satisfy Lintner's (L1) condition of a long-run target payout ratio where this ratio is given by $\delta\equiv\Sigma_0^N\lambda_k$.

Consider an economy in which the p^* test assumptions (S1) and (S2) *are known* to hold, but instead of (S3), assume that (9) describes the stochastic process for aggregate real dividends on the market portfolio. From the assumption of a constant discount rate (S2) and the definition of permanent earnings, we have from (11) that detrended real dividends at time t can be written as

$$(12)\qquad d(t)=r\sum_{k=0}^{N}\lambda_k v(t-k)$$

$$=r\delta\sum_{k=0}^{N}\theta_k v(t-k)$$

where $v(s)\equiv V(s)/(1+g)^s$ is the detrended real intrinsic value per share of the firm at time s and $\theta_k\equiv\lambda_k/\delta\geq 0$ with $\Sigma_0^N\theta_k=1$.

From (S1), stock prices are known to be rationally determined, and therefore, it follows from (7) that $p(t)=v(t)$ for all t. Hence, from (12), current detrended dividends can be expressed as a function of current and past detrended stock prices: namely,

$$(13)\qquad d(t)=\rho\sum_{k=0}^{N}\theta_k p(t-k),$$

where $\rho=r\delta$ is the long-run or steady-state dividend-to-price ratio on the market portfolio.[15] Thus, from (S1), (S2), and (9), detrended aggregate real dividends are a moving average of current and past detrended real stock prices. Moreover, under these posited conditions, the *ex post* rational price series constructed for the sample period $[0,T]$ can be expressed as a convex combination of the observed detrended stock prices, $p(t)$, $t=-N,\ldots,0,1,\ldots,T-1$. That is, from (3) and (13), (5) can be rewritten as

$$(14)\qquad p^*(t)=\sum_{k=-N}^{T-1}w_{tk}p(k),\qquad t=0,1,\ldots,T-1,$$

where, as can be easily shown, the derived weights satisfy

$$\sum_{k=-N}^{T-1}w_{tk}=1$$

and $w_{tk}\geq 0$ with $w_{tk}=0$ for $k<t-N$.

THEOREM 1: *If, for each* t, $p^*(t)=\Sigma_{k=0}^{T-1}\pi_{tk}p(k)$ *where* $\Sigma_{k=0}^{T-1}\pi_{tk}=1$; $\Sigma_{t=0}^{T-1}\pi_{tk}\leq 1$ *and* $\pi_{tk}\geq 0$, *then for each and every sample path of stock price realizations*, $\mathrm{Var}(p^*)\leq\mathrm{Var}(p)$, *with equality holding if and only if*

[14] The constant of integration must be zero since $e(t)=0$ implies that $V(t)=0$, which implies that $e(t+s)=0$ and $d(t+s)=0$ for all $s\geq 0$.

[15] The target payout ratio δ and the long-run growth rate g are related by $g=(1-\delta)r/[1+r\delta]$.

all realized prices are identical in the sample $t = 0, \ldots, T-1$.

The formal proof is in the Appendix. However, a brief intuitive explanation of the theorem is as follows: for each $t, t = 0, \ldots, T-1$, $p^*(t)$ is formally similar to a conditional expectation of a random variable p with possible outcomes $p(0), \ldots, p(T-1)$ where the $\{\pi_{tk}\}$ are interpreted as conditional probabilities. Var(p^*) is, therefore, similar to the variance of the conditional expectations of p which is always strictly less than the variance of p itself (unless, of course, Var(p) = 0).

The variance inequality in Theorem 1 is the exact opposite of inequality (4) which holds that $\text{Var}(p^*) \geq \text{Var}(p)$. That is, if the *ex post* rational price series satisfies the hypothesized conditions of Theorem 1, then the p^* test inequality will be violated whether or not actual stock prices are *ex ante* rational. Because Theorem 1 applies to each and every time path of prices, its derived inequality $\text{Var}(p^*) \leq \text{Var}(p)$ holds *in-sample*. A fortiori, it will obtain for any distribution of prices. Thus, even for a "bad draw," Var(p^*) will not exceed Var(p).

Although the inequality in Theorem 1 is an analytic result, it does not strictly hold for all possible sample paths of the $p^*(t)$ series generated by the dividend process (9) and rational stock prices. By inspection of (3) and (14), for each t, $N \leq t \leq T$, $p^*(t)$ is a convex combination of the sample stock prices $\{p(0), \ldots, p(T-1)\}$ that satisfies the hypothesized conditions of Theorem 1. However, for $0 \leq t \leq N-1$, $p^*(t)$ will depend upon both the sample period's stock prices and one or more "out-of-sample" stock prices $\{p(-N), \ldots, p(-1)\}$. Hence, with the exception of one member of the class of processes given by (9),[16] $\text{Var}(p^*) \leq \text{Var}(p)$ need not obtain for each and every sample path of prices.[17] The problem created here by out-of-sample prices is similar to the general "start-up" problem in using a finite sample to estimate a moving average or distributed lag process. Because only the first N of the T sample elements in the p^* series depend on out-of-sample prices, the influence of these prices on the sample variance of p^* becomes progressively smaller as the length of the sample period is increased. Indeed, as proved in the Appendix, we have that

THEOREM 2: *If* (S1) *and* (S2) *hold and if the process for aggregate real dividends is given by* (9), *then in the limit as* $T/N \to \infty$, $\text{Var}(p^*)/\text{Var}(p) \leq 1$ *will hold almost certainly.*

As noted in the introduction, the Shiller variance bound theorem has been widely interpreted as a test of stock market rationality. However, as with Theorem 1, Theorem 2 concludes that Var(p^*) is a *lower* bound on Var(p) whereas, the corresponding Shiller theorem concludes that Var(p^*) is an *upper* bound on Var(p). Both Theorem 2 and the Shiller theorem are mathematically correct and both share in common the hypothesis (S1) that stock prices are rationally determined. Therefore, if these variance bound theorems are interpreted as tests of stock market rationality, then we have the empirical paradox that this hypothesis can always be rejected. That is, if observed stock prices were to satisfy the p^* test of stock market rationality, then this same sample of prices must fail our test, and conversely. This finding alone casts considerable doubt on the reliability of such variance bound theorems as tests of stock market rationality.

The apparent empirical paradox is, of course, resolved by recognizing that each of the variance bound theorems provides a test of a different joint hypothesis. In addition to

[16] The exception is the polar case of (9) where $N = 0$ and mangers choose a dividend policy so as to maintain a target payout ratio in both the short and long run. In this case, with $d(t) = \rho p(t)$ for all t, we have the stronger analytic proposition that the Shiller variance bound inequality (4) must be violated in all samples if stock prices are rational.

[17] For example, if all in-sample prices happened to be the same (i.e., $p(t) = \bar{p}, t = 0, \ldots, T-1$), but the out-of-sample prices were not, then for that particular sample path, $\text{Var}(p^*) > \text{Var}(p) = 0$.

(S1), both theorems also assume that the real discount rate is constant. Hence, neither (S1) nor (S2) of the respective joint hypotheses is the source of each theorem's contradictory conclusion to the other.[18] It therefore follows necessarily that the class of aggregate dividend processes (9) postulated in Theorem 2 is incompatible with the Shiller theorem assumption (S3) of a regular stationary process for detrended aggregate dividends.[19] That is, given that (S1) and (S2) hold, nonstationarity of the dividend process is a necessary condition for the validity of Theorem 2[20] whereas stationarity of the dividend process is a sufficient condition for the validity of the p^* test inequality (4). Thus, the diametrically opposite conclusions of these variance bound theorems follow directly from the differences in their posited dividend processes.

In this light, it seems to us that if the p^* test is to be interpreted as a test of any single element of its joint hypothesis, (S1), (S2), and (S3), then it is more appropriately viewed as a test of (S3) than of (S1). Viewed in this way, the previously cited empirical findings of a large violation of inequality (4) would appear to provide a rather impressive rejection of the hypothesis that aggregate real dividends follow a stationary stochastic process with a trend. As noted, Shiller has argued extensively that his results are empirically robust with respect to assumption (S2). In a parallel fashion, we would argue that they are also robust with respect to (S1). That is, even if stock prices were irrationally volatile, the amount of irrationality required to "save" the stationarity hypothesis (S3) is so large that the measured five-to-thirteen times excess variation in stock prices cannot reasonably be attributed to this source.

Perhaps the p^* test might still be saved as a test of stock market rationality if there were compelling a priori economic reasons or empirical evidence to support a strong prior belief that aggregate dividends follow a stationary process with a trend. We are, however, unaware of any strong theoretical or empirical foundation for this belief. Indeed, the standard models in the theoretical and empirical literature of both financial economics and accounting assume that stock prices, earnings, and dividends are described by nonstationary processes.[21] In his analyses of the Shiller and other variance bounds tests, Kleidon (1983a, b) uses regression and other time-series methods to show that the hypothesis of stationarity for the aggregate Standard and Poor's 500 stock price, earnings, and dividend series can be rejected.

We (in our forthcoming article) develop and test an aggregate dividend model based on the same Lintner stylized facts used to motivate (9) here. In this model, the dividend-to-price ratio follows a stationary process, but both the dividend and stock price

[18]Since the two theorems share the assumption (S2) and for any sample of prices, one must fail, they cannot reliably be used to test this hypothesis either. However, as Eugene Fama (1977) and Stewart Myers and Stuart Turnbull (1977) have shown, we note that a constant discount rate is inconsistent with a stationary process for dividends when investors are risk averse. Hence, the assumptions (S2) and (S3) are a priori mutually inconsistent.

[19]If $V(t)$ follows a stationary process and the dividend process is given by (9), then the innovations or unanticipated changes in intrinsic value, $\Delta V(t) + D(t) - rV(t)$, will not form a martingale as is required by (6). If, as is necessary for the validity of (9), the intrinsic value follows a nonstationary process, then from (6) and (7), both dividends and rational stock prices must also be nonstationary.

[20]If $p(t)$ and $d(t)$ follow nonstationary processes, the variances of the price and dividend are, of course, not well-defined in the time-series sense that they were used in Shiller's variance bound test. However, $\mathrm{Var}(p^*)$ and $\mathrm{Var}(p)$ can be simply treated as sample statistics constructed from the random variables $\{p(t)\}$ and $\{d(t)\}$, and for any finite T, the conditional moments of their distributions will exist. If, moreover, the processes are such that the dividend-to-price ratio coverages to a finite-variance steady-state distribution, then the conditional expectation of the variance bound inequality as expressed in Theorem 2, $\varepsilon_0[\mathrm{Var}(p^*)/\mathrm{Var}(p)]$, will exist even in the limit as $T \to \infty$.

[21]In financial economics, the prototypical assumption is that the per period rates of return on stocks are independently and identically distributed over time. Together with limited liability on stock ownership, this implies a geometric Brownian motion model for stock prices which is not, of course, a stationary process. There is a long-standing and almost uniform agreement in the accounting literature that accounting earnings (either real or nominal) can best be described by a nonstationary process (see George Foster, 1978, ch. 4).

processes are themselves nonstationary. This model is shown to significantly outperform empirically the univariate autoregressive model (with a trend) normally associated with a stationary process. These results not only cast further doubt on the stationarity assumption, but also provide affirmative evidence in support of the class of dividend processes hypothesized in Theorems 1 and 2.

Our model can also be used to reinterpret other related empirical findings which purport to show that stock prices are too volatile. For example, to provide a more-visual (if less-quantitatively precise) representation of the "excess volatility" of stock prices, Shiller (1981a, p. 422) plots the time-series of the levels of actual detrended stock prices and the constructed *ex post* rational prices, $p^*(t)$. By inspection of these plots, it is readily apparent that $p(t)$ is more volatile than $p^*(t)$. Instead of implying "too much" stock price volatility, these plots can be interpreted as implying that the p^* series has "too little" volatility to be consistent with a dividend process which is not smoothed. They are, however, entirely consistent with rational and nonstationary stock prices and dividend policies like (9) which smooth the dividend process.

It also appears in these plots that the levels of actual prices "revert" toward the p^* trend line. In the context of (14), this apparent correspondence in trend should not be surprising since $p^*(t)$ is in effect a weighted sum of *future* actual prices that were, of course, not known to investors at time t. The *ex post* "mistakes" in forecasts of these future prices by the market at time t are, thus, "corrected" when the subsequent "right" prices (which were already contained in $p^*(t)$) are revealed.[22]

In his latest published remarks on the plots of these time-series, Shiller concludes:

> The near-total lack of correspondence, except for trend, between the aggregate stock price and its *ex post* rational counterpart (as shown in Figure 1 of my 1981a paper) means that essentially no observed movements in aggregate dividends were ever correctly forecast by movements in aggregate stock prices! [1983, p. 237]

This conclusion does not, however, appear to conform to the empirical facts. As shown in our forthcoming article, the single variable that provides, by far, the most significant and robust forecasting power of the subsequent year's change in aggregate dividends is the previous year's unanticipated change in aggregate stock price.[23]

Shiller (1981a, pp. 425–27) presents a second variance bound test of rational stock prices that uses the time-series of "price innovations" that he denotes by $\delta p(t) \equiv p(t)-p(t-1)+d(t-1)-\rho p(t-1)$. Under the assumption that detrended dividends have a stationary distribution, he derives as a condition for rational stock prices that

$$(15)\quad \mathrm{Var}(d) \geq \mathrm{Var}(\delta p)\left[(1+\rho)^2-1\right],$$

where $\mathrm{Var}(d)$ and $\mathrm{Var}(\delta p)$ denote the sample variances of the level of detrended dividends and the innovations of price changes, respectively. As reported in Shiller's cited Table 2, the null hypothesis of rational stock prices seems, once again, to be grossly violated by both his data sets.

If, however, dividends are generated by a process like (9) and rational stock prices follow a nonstationary process, then the inequality (15) is no longer valid. Moreover, in this case, it is likely that inequality (15) will be violated. Suppose, for example, that the innovations in (detrended) stock prices follow a geometric Brownian motion

[22] The strength of this apparent reversion to trend is further accentuated by using the *ex post* or in-sample trend of stock prices to detrend both the actual stock price and the $p^*(t)$ time-series.

[23] As shown in Fischer and Merton, in addition to predicting dividend changes, aggregate real stock price changes are among the better forecasters of future changes in business cycle variables including *GNP*, corporate earnings, and business fixed investment. These empirical findings might also be counted in the support of the hypothesis of stock market rationality.

given by

$$(16)\qquad \delta p(t) = \sigma p(t-1)Z(t),$$

where $\{Z(t)\}$ are independently and identically distributed random variables with $\varepsilon_{t-1}[Z(t)]=0$; $\varepsilon_{t-1}[Z^2(t)]=1$; σ, a positive constant and where ε_t is the expectation operator, conditional on information available at time t. It follows from (16) and the properties of $\{Z(t)\}$ that

$$(17)\quad \varepsilon_0[\mathrm{Var}(\delta p)] = \varepsilon_0\left[\sum_{t=0}^{T-1}\sigma^2 p^2(t)\right]\Big/T,$$

$$= \sigma^2\varepsilon_0\left[\mathrm{Var}(p)+(\bar{p})^2\right],$$

where $\bar{p} = p^*(T)$ given in (3).

From the posited dividend process given in (13), $d(t)/\rho$ is a distributed lag of past stock prices where the distribution weights are nonnegative and sum to unity. Thus, the ensemble $\{d(t)/\rho\}$ satisfies the hypothesized conditions of Theorems 1 and 2. It follows, therefore, that $\mathrm{Var}(d/\rho) \le \mathrm{Var}(p)$. Factoring out the constant ρ and rearranging terms, the inequality can be rewritten as

$$(18)\qquad \mathrm{Var}(d) \le \rho^2\,\mathrm{Var}(p),$$

with equality holding only in the special limiting case of (13) with $d(t) = \rho p(t)$. Combining (17) and (18) and rearranging terms, we have that

(19)

$$\varepsilon_0[\mathrm{Var}(d)] \le \rho^2\varepsilon_0\left[\mathrm{Var}(\delta p) - \sigma^2\bar{p}^2\right]/\sigma^2.$$

In sharp contrast to the stationary dividend case in (15) where the variance of dividends provides an upper bound on the volatility of rational stock price innovations, inspection of (19) shows that this variance provides only a *lower* bound on that volatility for our dividend process.[24] Inequalities (15) and (19) are not, of course, mutually exclusive for all parameter values. However, using the estimated values of $\mathrm{Var}(\delta p)$, $\bar{p}$, $\mathrm{Var}(p)$, and ρ reported by Shiller for his 1871–1979 Standard and Poor's data set, and our equation (17), we have that $\hat{\rho} = 0.0480$ and $\hat{\sigma}^2 = \mathrm{Var}(\delta p)/[\mathrm{Var}(p)+\bar{p}^2] = 0.0276$. Substitution of these values in (15) implies that $\varepsilon_0[\mathrm{Var}(d)] \ge .0983\,\varepsilon_0[\mathrm{Var}(\delta p)]$ whereas the same values substituted in (19) implies that $\varepsilon_0[\mathrm{Var}(d)] \le .0835\,\varepsilon_0[\mathrm{Var}(\delta p) - .0276(\bar{p})^2]$. Thus, given these parameter values it would appear that any recorded values for $\mathrm{Var}(\delta p)$ and $\mathrm{Var}(d)$ would violate one or the other variance bound inequalities for rational stock price innovations. Hence, the empirical finding that $\mathrm{Var}(d) \ll .0983\,\mathrm{Var}(\delta p)$—although inconsistent with the stationarity assumption (S3)— is entirely consistent with rational stock prices and the aggregate dividend process (9).

We are not alone in questioning the specification of the dividend process in the Shiller model. In addition to the cited Kleidon analyses, Basil Copeland (1983) has commented on the assumption of a deterministic trend. In his reply to Copeland, Shiller had this to say on the specification issue:[25] "Of course, we do not literally believe with certainty all the assumptions in the model which are the basis of testing. I did not intend to assert in the paper that I know dividends were indeed stationary around the historical trend" (1983, p. 236). We have shown, however, that variance bound inequality (4) is critically sensitive to the assumption of a stationary process for aggregate dividends. If aggregate dividend policy is described by a smoothing or averaging of intrinsic values that follow a nonstationary process, then the misspecification of stationarity in the dividend process does not

[24] By inspection of (18) and (19), it is evident that the strength of the inequality (19) will depend on the degree of "dividend smoothing" undertaken by managers. Generally, the larger is N and the more evenly distributed the weights $\{\theta_k\}$ in (13), the smaller will be $\mathrm{Var}(d)$ with no corresponding reduction in $\mathrm{Var}(\delta p)$.

[25] We surely echo this view with respect to our own dividend model (9). We do not however, assert that the variance bound condition of Theorems 1 and 2 provides a reliable method for testing stock market rationality.

just weaken the power of this bound as a test of stock market rationality—it destroys it—because in that case the fundamental inequality is exactly reversed.

In summary, the story that dividends follow a stationary process with a trend leads to the empirical conclusion that aggregate stock prices are *grossly* irrational. It has, therefore, the deep and wide-ranging implications for economic theory and policy that follow from this conclusion. The majority of empirical tests of the efficient market theory do not, however, concur with this finding. Hence, to accept this dividend story, we must further conclude that the methodologies of these tests were sufficiently flawed that they failed to reject this hypothesis in spite of the implied substantial irrationality in stock prices. Similar flaws must also be ascribed to the extensive studies in finance and accounting that claim to show earnings, dividends, and stock prices follow nonstationary processes. If, however, this dividend story is rejected, then the empirical violation of inequalities (4) and (15) implies nothing at all about stock market rationality. In the spirit of Edward Leamer's (1983) discussion of hypothesis testing, we therefore conclude that the Shiller variance bound theorem is a wholly unreliable test of stock market rationality because, as Leamer said, "...there are assumptions within the set under consideration that lead to radically different conclusions" (p. 38).

II. Overview and Conclusion

In the previously cited reply to Copeland, Shiller proclaims:

> The challenge for advocates of the efficient markets model is to tell a convincing story which is consistent both with observed trendiness of dividends for a century and with the high volatility of stock prices. They can certainly tell a story which is within the realm of possibility, but it is hard to see how they could come up with the inspiring evidence for the model. [1983, p. 237]

We believe that the theoretical and empirical analysis presented here provides such "inspiring evidence."

In general, the statistical properties of estimators drawn from a nonstationary population are an important matter in evaluating the significance of variance bound inequality violations.[26] As it happens, our reconciliation of the gross empirical violations of Shiller's variance bound inequalities is not based on sampling arguments. That is, we do not merely show that it is *possible* to have a chance run of history where the measured volatility of dividends is greatly exceeded by the measured volatility of rational stock prices. Instead, our "reversals" of variance bound inequalities (4) and (15) are based on expected values over the population. Thus, over repeated runs of history, we expect that, on average, these inequalities would be violated.

If our results seem counterintuitive to some, then perhaps this indicates that such intuitions about volatility relations between optimal forecasts and realizations are implicitly based on the assumption of stationary and linear processes. If so, we hope that this analysis serves to illustrate the potential for cognitive misperceptions from applying such intuitions to nonstationary systems.

Economists have long known that fluctuations in stock prices are considerably larger than the fluctuations in aggregate consumption, national income, the money supply, and many other similar variables whose expected future values presumably play a part in the rational determination of stock prices. Indeed, as noted in the introduction, we suspect that the sympathetic view held by some economists toward the proposition of excess stock market volatility can largely be traced to this long-established observation. Those who make this inference implicitly assume that the level of variability observed in these economic variables provides the appropriate

[26] Kenneth West (1984) and N. Gregory Mankiw et al. (1985) derive variance bound inequalities which appear to apply to nonstationary processes for prices and dividends. Using the Shiller data, Mankiw et al. find their inequalities are grossly violated. By studying the statistical properties of their estimates, Merton (1986) shows that the Mankiw et al. tests cannot be reliably applied when stock prices follow diffusion processes such as the geometric Brownian motion.

frame of reference from which to judge the rationality of observed stock price volatility. Although quantitatively more precise, the Shiller analysis adopts this same perspective when it asks: "If stock prices are rational, then why are they so volatile (relative to dividends)?" The apparent answer is that stock prices are not rational.

Our analysis turns this perspective "on its head" by asking: "If stock prices are rational, then why do dividends exhibit so little volatility (relative to stock prices)?" Our answer is simply that managers choose dividend policies so as to smooth the effect of changes in intrinsic values (and hence, rational stock prices) on the change in dividends. The a priori economic arguments and empirical support presented for this conclusion surely need no repeating. We would note, however, that this explanation is likely to also apply to the time-series of other economic flow variables. There are, for example, good economic reasons for believing that aggregate accounting earnings, investment, and consumption have in common with dividends that their changes are smoothed either by the behavior of the economic agents that control them or by the statistical methods which are used to measure them. An initial examination of the data appears to support this belief. If a thorough empirical evaluation confirms this finding, then our analysis casts doubt in general over the use of volatility comparisons between stock prices and economic variables which are not also speculative prices, as a methodology to test stock market rationality.

In summary of our view of the current state of the debate over the efficient market theory, Samuelson said it well when he addressed the practicing investment managers of the financial community over a decade ago:

> Indeed, to reveal my bias, the ball is in the court of the practical men: it is the turn of the Mountain to take a first step toward the theoretical Mohammed....
>
> ...If you oversimplify the debate, it can be put in the form of the question,
>
> Resolved, that the best of money managers cannot be demonstrated to be able to deliver the goods of superior portfolio-selection performance.
>
> Any jury that reviews the evidence, and there is a great deal of relevant evidence, must at least come out with the Scottish verdict:
>
> Superior Investment performance is unproved. [1974, pp. 18–19]

Just so, our evidence does not prove that the market is efficient, but it does at least warrant the Scottish verdict:

Excess stock price volatility is unproved. The ball is once again in the court of those who doubt the Efficient Market Hypothesis.

Appendix

PROOF of Theorem 1:

Define Π as the $T\times K$ matrix of elements π_{tk} in Theorem 1, so that $\boldsymbol{p}^*=\Pi\boldsymbol{p}$. We show that the following three conditions

$$\text{(A1)}\qquad \pi_{tk}>0$$

$$\text{(A2)}\qquad \sum_{k=1}^{T}\pi_{tk}=1 \qquad \text{for all} \quad t=1,\dots,T$$

$$\text{(A3)}\qquad \sum_{t=1}^{T}\pi_{tk}\leq 1 \qquad \text{for all} \quad k=1,\dots,T$$

are sufficient for

$$\text{(A4)}\qquad \operatorname{Var}(\Pi\boldsymbol{p})\leq\operatorname{Var}(\boldsymbol{p}),$$

where $\operatorname{Var}(\boldsymbol{x})$ is defined as the *sample* variance operator applied to the elements of $\boldsymbol{x}$.

LEMMA: *Any given* $\mathbf{p}$ *can be decomposed as* $\bar{p}\iota+\tilde{p}$ *where* $\bar{p}$ *is the sample mean,* ι *is a vector of ones, and* $\tilde{\mathbf{p}}$ *is the vector of deviations of the elements of* $\mathbf{p}$ *about* $\bar{p}$. *Then,*

$$\text{(A5)}\qquad \tilde{\boldsymbol{p}}'\Pi'\left(I-\frac{1}{T}\iota\iota'\right)\Pi\tilde{\boldsymbol{p}}\leq\tilde{\boldsymbol{p}}'\tilde{\boldsymbol{p}}$$

implies

$$\text{(A6)}\quad \operatorname{Var}(\Pi \boldsymbol{p}) = \left[\frac{\boldsymbol{p}'\Pi'\left(I - \frac{1}{T}\iota\iota'\right)\Pi\boldsymbol{p}}{T}\right] \le \left[\frac{\boldsymbol{p}'\left(I - \frac{1}{T}\iota\iota'\right)\boldsymbol{p}}{T}\right] = \operatorname{Var}(\boldsymbol{p})$$

under condition (A2) *where* $\operatorname{Var}(\Pi\boldsymbol{p})$ *is computed with respect to the T elements of* $\Pi\boldsymbol{p}$, *and* $\operatorname{Var}(\boldsymbol{p})$ *is defined with respect to the T elements of* $\boldsymbol{p}$. *If the matrix* Π *is rectangular of dimension* $T \times X$, *then the additional constraint* $\sum_{t=1}^{T}\pi_{tk} \le T/K$ *is sufficient for equation* (A6).

PROOF:

Substitute the decomposition $\boldsymbol{p} = \bar{p}\iota + \tilde{\boldsymbol{p}}$ into (A5) and realize that $(I - 1/T\iota\iota')\bar{p}\iota = 0$ and $(I - 1/T\iota\iota')\Pi\bar{p}\iota = 0$.

To prove Theorem 1, define the norm:

$$\text{(A7)}\qquad \|\Pi\tilde{\boldsymbol{p}}\|^2 = \sum_t \left(\sum_k \pi_{tk}\tilde{p}_k\right)^2.$$

Since the function $f(u) = u^2$ is convex, it follows that if $p_{tk} > 0$ and $\Sigma_k \pi_{tk} = 1$, then,

$$\text{(A8)}\qquad \left(\sum_k \pi_{tk}\tilde{p}_k\right)^2 \le \sum_k \pi_{tk}\tilde{p}_k^2,$$

so

$$\text{(A9)}\quad \|\Pi\tilde{\boldsymbol{p}}\|^2 \le \sum_t\sum_k \pi_{tk}\tilde{p}_k^2 = \sum_k \tilde{p}_k^2 \sum_t \pi_{tk} \le \sum_k \tilde{p}_k^2.$$

(The last inequality in (A10) is strict if $\Sigma_t \pi_{tk} < 1$). Equation (A9) can be rewritten as

$$\text{(A10)}\qquad \|\Pi\tilde{\boldsymbol{p}}\| \le \|\tilde{\boldsymbol{p}}\|.$$

Also,

$$\text{(A11)}\qquad \tilde{\boldsymbol{p}}'\Pi'\left(I - \frac{1}{T}\iota\iota'\right)\Pi\tilde{\boldsymbol{p}} \le \|\Pi\tilde{\boldsymbol{p}}\|.$$

(A10) and (A11) together imply (A5) which, by the Lemma, implies (A6), that is,

$$\text{(A12)}\qquad \operatorname{Var}(\Pi\boldsymbol{p}) \le \operatorname{Var}(\boldsymbol{p}).$$

PROOF of Theorem 2:

Using the definition of the (detrended) *ex post* rational price, $p^*(t)$, given in (5), and allowing (detrended) dividends to be a general distributed lag of (detrended) prices as in (13), *ex post* rational prices can be expressed in terms of the observed and presample (detrended) prices as

$$\text{(A13)}\quad \begin{bmatrix} p^*(T-1)\,p^*(T-2) \\ p(T-3) \\ \vdots \\ p^*(1) \end{bmatrix} = \left\{ \frac{1}{T}\begin{bmatrix} \eta\ldots\eta & 0\ldots0 \\ \eta^2\ldots\eta^2 & 0\ldots0 \\ \eta^3\ldots\eta^3 & 0\ldots0 \\ \vdots\quad\vdots & \\ \eta^T\ldots\eta^T & 0\ldots0 \end{bmatrix} + \rho\begin{bmatrix} \eta & 0 & 0\ldots0 \\ \eta^2 & \eta & 0\ldots0 \\ \eta^3 & \eta^2 & \eta\ldots0 \\ \vdots & \vdots & \vdots \\ \eta^T & \eta^{T-1} & \eta^{T-2}\ldots0 \end{bmatrix} \times \begin{bmatrix} \theta_0 & \theta_1\ldots\theta_N & 0\ldots0 & 0\ldots0 \\ 0 & \theta_0\ldots\theta_{N-1} & \theta_N\ \ 0 & 0\ldots0 \\ \vdots & \vdots & \vdots & \vdots\ \ \vdots \\ \vdots & \vdots & \theta_N & 0\ \ 0 \\ \vdots & \vdots & \vdots & \vdots\ \ \vdots \\ 0 & 0 & \theta_0\theta_1 & 0\ldots0 \\ 0 & 0\ldots0 & 0\ldots\theta_0 & \theta_1\ldots\theta_N \end{bmatrix} \right\} \times \begin{bmatrix} p(T-1) \\ p(T-2) \\ \vdots \\ p(1) \\ p(-1) \\ \vdots \\ p(-N) \end{bmatrix}$$

where $\eta \equiv (1+g)/(1+r) \equiv 1/(1+r\delta) \equiv 1/(1+\rho)$ (the first identity follows from the definition of η in (1), the second from fn. 15, and the third from the definition of ρ in (15)) and the level of dividends is a distributed lag of the level of past prices, as in (13), that is, $d(t) = \rho\Sigma_{k=0}^{N}\theta_k p(t-k)$.

Equation (A13) may be conveniently rewritten as

$$\text{(A14)} \qquad \boldsymbol{p}^* = [A_1 + A_2\Theta]\boldsymbol{p}$$

where A_1 is the first matrix on the right-hand side of (A13), that is, the matrix that involves multiplication by the scalar $1/T$), A_2 is the next matrix, that involves multiplication by the scalar ρ, and Θ is the matrix that contains the elements $\theta_1, \theta_2, \ldots, \theta_N$. The weights in the matrix A_1 reflect the contribution $1/T$ of each of the observed prices $[p(T-1), \ldots p(0)]$ to $p^*(T)$ in accordance with (3), together with the weight $[1/(1+\rho)]^{T-t}$ attached to $p^*(T)$ in the determination of $p^*(t)$ in (5). The matrix A_2 contains the discount weights that (5) places on dividends as components of each $p^*(t)$, while Θ contains the distributed lag weights of dividends on past prices, as given in (13). Using these definitions of A_1, A_2, and Θ, (A14) is equivalent to

$$\text{(A15)} \qquad \boldsymbol{p}^* = W\boldsymbol{p},$$

where $W = [w_{tk}]$, the w_{tk} being those defined in (14).

It may be verified that the component A_1 of the transformation matrix W is irrelevant to the application of Theorem 1 (because the proof proceeds in terms of $\tilde{\boldsymbol{p}}$, the deviations of the elements of $\boldsymbol{p}$ about $\bar{p}$). The elements of $A_2\Theta$ are positive and sum to unity or less across the rows, and if $\Theta = I$, or Θ is such that column elements of $A_2\Theta$ sum to less that $T/(T+N)$ if $\Theta \neq I$, then the conditions of Theorem 1 are satisfied, and we have

$$\text{(A16)} \qquad \operatorname{Var}(\boldsymbol{p}^*) \leq \operatorname{Var}(\boldsymbol{p}).$$

In the market rationality tests, the variance of the *ex post* rational prices is compared not to the variance of the $(T+N)$ vector of T observed *and* N presample prices, but to the variance of only the T observed prices. Partitioning of the $(T+N)$ prices into in-sample and out-of-sample prices, it is straightforward to show that

$$\text{(A17)} \quad \operatorname{Var}(p^*)/\operatorname{Var}(p_T) \leq 1 + \frac{T.N}{(N+T)^2}\frac{(\bar{p}_T-\bar{p}_N)^2}{\operatorname{Var}(p_T)} - \frac{N}{(N+T)}\left[\frac{\operatorname{Var}(p_T)-\operatorname{Var}(p_N)}{\operatorname{Var}(p_T)}\right],$$

where $\bar{p}_T = \sum_{t=0}^{T-1} p(t)/T$;

$$\bar{p}_N = \sum_{t=-N}^{-1} p(t)/N;$$

$$\operatorname{Var}(p_T) = \sum_{t=0}^{T-1}[p(t)-\bar{p}_T]^2/T;$$

$$\operatorname{Var}(p_N) = \sum_{t=-N}^{-1}[p(t)-\bar{p}_N]^2/N.$$

The sum of the last two terms on the right-hand side of (A17) can be positive for some sample paths. However, if N is finite, and the nonstationary process for prices is not degenerate, then it is clear that the start-up adjustment terms in (A17) converge in mean square to zero as $T \to \infty$.

REFERENCES

Bosworth, Barry, "The Stock Market and the Economy," *Brookings Papers on Economic Activity*, 2:1975, 257–300.

Copeland, Basil L., Jr., "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?: Comment," *American Economic Review*, March 1983, 73, 234–35.

Fama, Eugene F., "Risk-Adjusted Discount Rates and Capital Budgeting under Uncertainty," *Journal of Financial Economics*, August 1977, *5*, 3–24.

Fischer, Stanley and Merton, Robert C., "Mac-

roeconomics and Finance: The Role of the Stock Market," *Carnegie-Rochester Conference Series on Public Policy*, Vol. 21, *Essays on Macroeconomic Implications of Financial and Labor Markets and Political Processes*, Autumn 1984, 57–108.

Flavin, Marjorie A., "Excess Volatility in the Financial Markets: A Reassessment of the Empirical Evidence," *Journal of Political Economy*, December 1983, *91*, 929–56.

Foster, George, *Financial Statement Analysis*, Englewood Cliffs: Prentice-Hall, 1978.

Keynes, John Maynard, *The General Theory of Employment Interest and Money*, New York: Harcourt, Brace, 1936.

Kleidon, Allan W., (1983a) "Variance Bounds Tests and Stock Price Valuation Models," unpublished working paper, Graduate School of Business, Stanford University, January 1983.

______, (1983b) "Bias in Small Sample Tests of Stock Price Rationality," unpublished paper, University of Chicago, 1983.

Leamer, Edward E., "Let's Take the Con Out of Econometrics," *American Economic Review*, March 1983, *73*, 31–43.

LeRoy, Stephen F. and Porter, Richard D., "The Present-Value Relation: Tests Based on Implied Variance Bounds," *Econometrica*, May 1981, *49*, 555–74.

Lintner, John, "Distribution of Incomes of Corporations Among Dividends, Retained Earnings, and Taxes," *American Economic Review Proceedings*, May 1956, *66*, 97–113.

Mankiw, N. Gregory, Romer, David and Shapiro, Matthew D., "An Unbiased Reexamination of Stock Market Volatility, *Journal of Finance*, July 1985, *40*, 677–78.

Marsh, Terry A. and Merton, Robert C., "Aggregate Dividend Behavior and Its implications for Tests of Stock Market Rationality," Working Paper 1475-83, Sloan School of Management, September 1983.

______ **and** ______, "Earnings Variability and Variance Bounds Tests for the Rationality of Stock Market Prices," Working Paper 1559-84, Sloan School of Management, March 1984.

______ **and** ______, "Dividend Behavior for the Aggregate Stock Market," *Journal of Business*, forthcoming.

Merton, Robert C., "On the Current State of the Stock Market Rationality Hypothesis," in Stanley Fischer et al., eds., *Macroeconomics and Finance: Essays in Honor of Franco Modigliani*, Cambridge: MIT Press, 1986.

Miller, Merton H. and Modigliani, Franco, "Dividend Policy, Growth and the Valuation of Shares," *Journal of Business*, October 1961, *34*, 411–33.

Myers, Stewart C. and Turnbull, Stuart M., "Capital Budgeting and the Capital Asset Pricing Model: Good News and Bad News," *Journal of Finance*, May 1977, *32*, 321–33.

Samuelson, Paul A., "Proof That Properly Anticipated Prices Fluctuate Randomly," *Industrial Management Review*, Spring 1965, *6*, 41–49.

______, "Challenge to Judgment," *Journal of Portfolio Management*, Fall 1974, *1*, 17–19.

______, "Optimality of Sluggish Predictors Under Ergodic Probabilities, *International Economic Review*, February 1976, *17*, 1–7.

Shiller, Robert J., (1981a) "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?," *American Economic Review*, June 1981, *71*, 421–36.

______, (1981b) "The Use of Volatility Measures in Assessing Market Efficiency," *Journal of Finance*, May 1981, *36*, 291–311.

______, "Consumption, Asset Markets, and Macroeconomic Fluctuations," *Carnegie-Rochester Conference Series on Public Policy*, Vol. 17, *Essays on Economic Policy in a World of Change*, 1982, 203–38.

______, "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in dividends?: Reply," *American Economic Review*, March 1983, *73*, 236–37.

Tobin, James, "On the Efficiency of the Financial System," Hirsch Memorial Lecture, New York, May 15, 1984.

West, Kenneth D., "Speculative Bubbles and Stock Price Volatility," Financial Research Center, Memo. No. 54, Princeton University, December 1984.

The Marsh-Merton Model of Managers' Smoothing of Dividends

By ROBERT J. SHILLER*

The fact that firms seem to follow an earnings payout policy that results in a "smoothed" dividend stream has often been brought up in criticism of the variance inequality tests that I used (1981a) to call into question the simple efficient markets model. It seems that the smoothing lowers the variance of detrended real dividends, and this may account for the apparent inadequacy of dividends movements to account for price movements. The fact that the present value of actual dividends p^* is itself a moving average of dividends d, and hence a smoothed version of dividends, is also brought up in criticism of the inequality that involves p^*. However, dividend smoothing or the smoothing implicit in p^* does not pose any problems for the theoretical volatility inequalities. As long as (real detrended) price p is the present value of expected (real detrended) dividends d, then the dividends, whether they are smoothed or not, must move enough according to the measures in the inequalities if the price movements are to be justified.

Terry Marsh and Robert Merton (1986) use arguments relying on the above-noted smoothing relations to show a sense in which, for sample variances, the variance inequalities in my paper may be thought of as reversed. Marsh and Merton model the behavior of those decision makers who set the level of dividends; the model (13) in their paper is a dividend smoothing model. Moreover, the proof of their Theorem 2 also makes use of the smoothing implicit in p^*. However, the feature of the model that causes the variance inequalities to be invalidated is not the smoothing per se, but the nonstationarity in dividends that is induced by the particular dividend smoothing rule (13).

Substituting their equation (13) into their equation (2) and then into their equation (1), we find

$$(1)\quad p(t) = \rho E_t \sum_{j=0}^{\infty} \eta^{j+1} \sum_{k=0}^{N} (\theta(k) p(t+j-k).$$

This is a rational expectations model whose characteristic polynomial has a unit root —since the right-hand side coefficients sum to one. Stochastic processes $p(t)$ that satisfy it will show no mean-reverting tendencies. Adding a constant to $p(t)$ for all t will not change the path of changes in p, since the constant would drop out of the above equation. This is so whether or not the $\theta(k)$ are all positive (so that dividends are smoothed). The nonstationarity would occur as well for any pattern of $\theta(k)$. (That $\Sigma\theta(k) = 1$ is not an assumption with content but restricts (13) as a result of the way detrending is done.) When the population variance is infinite, the sample variance (which is of course always finite) is necessarily a downward biassed measure of the population variance which enters the theoretical inequalities.

It is easiest to see what their model is doing by considering the special case $N=0$, so that $d(t) = \delta rp(t)$. A solution to the above rational expectations model then occurs if price $p(t)$ and dividend $d(t)$ are random walks. Now the sample variance over an interval of time of a random walk or the moving average of a random walk tends to increase without bound as the length of the sample interval is increased. It is thus possible that the efficient markets model could be true and still the variation in the dividend in any sample would appear inadequate (rela-

*Yale University, New Haven CT 06520 and National Bureau of Economic Research.

tive to the inequality $\mathrm{Var}(p^*) \geq \mathrm{Var}(p)$) to justify the variation in price. Essentially, the variation in price over the sample could be justified by looking at the higher variance of dividends as measured in terms of $\mathrm{Var}(p^*)$ over a longer sample period. But, of course, if one computed the variance in price p over the longer sample period as well, one would find that it was increased, too, so that the problem of apparent inadequacy of dividend variability would reappear.

This possibility would not appear relevant to understanding the behavior of the actual dividend and price series that we observe. If one looks at a plot of detrended dividends (see my 1981b article) one will see that it appears to be quickly and repeatedly mean reverting. If the trend of dividends is estimated (by regressing log dividend on time) separately for the two halves of the sample (1871 to 1924 and 1925 to 1978), the estimated coefficients of time for the two halves are almost identical: 1.18 and 1.33 percent. The standard deviations of the (separately) detrended dividends in the two halves are 1.07 and 1.41, respectively, which compare nicely with 1.27 for the whole sample.

Of course, the nonstationarity of dividends does not necessarily have to be obvious from the sample for it to justify the violation of the variance inequalities in the sample. Marsh and Merton could be right in their model of dividend setting. To say they are clearly wrong in their interpretation of the variance inequalities would require that their model be rejected at high confidence levels. A simple Dickey-Fuller test was presented in my 1981b article that rejected at the 5 percent level a random walk model of log dividends in favor of a model of stationary movements around a trend.[1] However, that evidence does not tell us whether a more complicated model along the lines of the general model in the Marsh-Merton paper could be rejected.[2] More generally, it seems intuitively implausible that estimating arbitrary time-series models within the sample will conclusively answer the fundamental question whether dividends are potentially much more variable than they were in the sample.

Marsh and Merton, incidentally, put no error term in (13) and so the model can be rejected with certainty. If they had included an error term, then the senses in which they see reversals of the variance inequalities would not be exactly right, but probably something of their spirit could be correct.

Of course, the stationarity assumption that I made in my earlier papers can't be exactly right either. One does not expect that the dividend process will stay right along the same exponential trend growth path forever. But, as with all modeling, the hope is that the model is sufficiently close to reality that the tests based on it are not fundamentally misleading. Indeed, some models of nonstationary dividends will give approximately the same volatility tests. The question is whether the Marsh-Merton story is sufficiently compelling that we think that the potential variability of dividends suggested by their model is a better measure of the true variability than is the historically observed variability around a trend.

The possibility that sample variances might not properly measure the population variances has been discussed a lot, even in my original paper (1981a) and that of Stephen LeRoy and Richard Porter (1981), but the specific issues that have been brought up are not usually those in the Marsh-Merton paper. Issues that have been tossed out as possibilities are such potential events that might dramatically change dividends in the future as oil crises, changes in tax laws, law suits, or nationalizations. In the Marsh-Merton

[1] Allan Kleidon (1986) did such a test using only data since 1926 and was unable to reject the random walk model at the 10 percent level. Throwing away half of the data (which he justified on the grounds of possible data errors in the earlier data) of course loses power. The data error problem is, incidentally, one against which the variance inequalities were meant to be robust; it is unfortunate that the problem is being reintroduced for the objective of evaluating stationarity.

[2] G. William Schwert (1985) has warned of the dangers of using these tests for unit roots when in fact the true process has a moving average component.

paper, all that is modeled is the rule-of-thumb dividend practice of managers, as a feedback on past prices of shares. In each period their own dividend-setting behavior affects price and, then in the next time period, they respond mechanically to the changed price as if they didn't realize that they themselves had changed it in the preceding period. The nonstationarity of dividends arises because of this feedback mechanism. No reason is offered why managers ought to be expected to do this: the model is justified only as a possible example of what Merton Miller (1977) called a "neutral mutation." Yet the theory that dividend payouts have no economic significance is belied by the fact that stock prices do move in response to dividend announcements.

It's a good idea to be a little suspicious of the use Marsh and Merton made of their simple model of dividend-smoothing behavior; that is, to demonstrate the importance of dividend nonstationarity. The basic idea of the model was purportedly envisioned by John Lintner (1956) as he sought to quantify what he had learned in interviews with those who decide on dividends. Interviews are useful adjuncts to efforts to model of human behavior, and I have used such interviews myself for this purpose. But we must bear in mind how vulnerable to error such methods are, and how we must work at the process of quantifying what we have learned from interviews. The Lintner model, as he wrote it down, worked out to be a Koyck-distributed lag model (the fashionable new econometric model in the mid-1950's) relating two readily available series: nominal dividends and nominal earnings. He did experiment with other formulations, but it is natural to suspect that the model might have preceded the interviews at least in part. The model may not be robust enough to allow consideration of the basic stationarity issue central to this paper.

Earning retentions are modeled by Lintner and Marsh and Merton as having nothing to do with internal investment opportunities of the firm. Yet Lintner noted (p. 104) that in his interviews such opportunities were among the "important factors" that managers discussed in connection with their dividend payout decisions. Moreover, investment opportunities have figured largely in other discussions of payout ratios (see, for example, Stewart Myers, 1984). Retained earnings are in fact used primarily for internal investment in the firm; rarely have they been used to buy back shares. If firms are more or less restricted to internal investments, then their growth is limited by the growth allowed by these opportunities.

The Marsh-Merton model (13) is not the same as Lintner's, in that "permanent earnings" (i.e., price) replaces earnings, and real rather than nominal values are used. It's worthwhile, then, to check its fit. I regressed (using the Standard and Poor data 1891–1979 from my 1981a paper) d on a constant, time, and p with 21-year (for $N=20$) second-degree polynomial distributed lag on p. The first distributed lag coefficient was freed from the polynomial constraint and a Cochrane-Orcutt serial correlation correction was used. The R-squared (squared correlation between d and its predicted value) was only .70. Both time and a constant term were omitted from the Marsh-Merton model: in this regression time was significant at the 10 percent level and the constant was significant at the 5 percent level. Time had a small negative coefficient and the constant term was positive. Most of the explanatory power comes from the contemporaneous value of p. These results might be construed as not necessarily contradicting the spirit of the Marsh-Merton analysis, but contradicting their specific claims for reversal of inequalities.

Even if we suppose that the Marsh and Merton analysis is basically sound, it may still be that the large violation of the variance inequalities that were observed may reflect some problem with the model. Marsh and Merton claim to give reasons to expect that $\sigma(p) \geq \sigma(p^*)$ in the sample, but the actual sample standard deviation of p was 5.6 times greater than that of p^*. Recent papers by Marjorie Flavin (1983) and Allan Kleidon (1986) show examples of stochastic processes in which violation of the variance inequalities is probable but extreme violations of them are improbable. Kleidon, for example, found that even in a random walk

case for log dividends the probability that $\sigma(p)/\sigma(p^*)$ is greater than 5 in a sample of 100 observations with a discount rate r of .065 (or very nearly the .0636 used in the computation of p^* in my 1981a paper) is only .148.[3]

It's important, however, not to dwell too long on the unit roots problem as it affected early tests of excess volatility of stock prices. Other ways of detrending stock prices that may deal satisfactorily with such nonstationarity are in the literature. Gregory Mankiw et al. (1985) noted that if we define $P^0(t)$ as any variable in the public information set at time t, and writing $P^*(t)-P^0(t)=(P^*(t)-P(t))+(P(t)-P^0(t))$, then the efficient markets model implies that the two expressions in parentheses on the right-hand side must be uncorrelated. Thus:

$$\text{(2)} \qquad \operatorname{Var}(P^*-P^0)\geq \operatorname{Var}(P^*-P),$$

$$\text{(3)} \qquad \operatorname{Var}(P^*-P^0)\geq \operatorname{Var}(P-P^0).$$

Kenneth West (1984) showed that the variance of the innovation $\delta P(t)=P(t)+D(t-1)-(1+r)P(t-1)$ must be less than or equal to the variance of the one-period forecast error for P^* using a forecasting model based on any subset of the public information set. (John Campbell, in unpublished work, showed that the West inequality follows from the Mankiw et al. inequalities using the fact that $P^*(t)-P(t)=\Sigma(k=1,\infty)\delta P(t+k)/(1+r)^k$.) Both Mankiw et al. (using $P_t^0=D^t/r$) and West (using an autoregressive model for the first difference of dividends) found violations of their variance inequalities, and their results are not vulnerable to criticisms that the series should be differenced to induce stationarity. Campbell and I (1985), using a variance inequality analogous to West's, found evidence of excess volatility in the context of a cointegrated model of D and P.[4] All of these violations of the variance inequalities were less dramatic than in my original paper.

It should be noted that it is possible that one may save the general notion of efficient markets by a route very different from that taken by Marsh and Merton, by attributing the variability of stock prices to changes in the discount rate rather than to the variability of expectations of future dividends. However, if changing discount rates were to be responsible for the major movements in stock prices, then one would expect that there would be substantial comovements between the stock market and the prices of long-term bonds, land, and housing. Such comovements do not seem to be in evidence. I still think that there may be an element of truth to such alternatives, and that more work on models along these lines may be a route toward bringing this out.

We should, on the other hand, not put too much hope for optimizing models of any sort in the explanation of stock price fluctuations. What we know about human psychology, changes in attitudes, fashions and fads, suggests otherwise. There is really absolutely no reason to think that movements in the aggregate stock market can be interpreted in terms of fundamentals. There is unfortunately a widespread misconception that any role for changing fashions or the like in determining financial asset prices has been soundly disproved by many statistical tests of market efficiency. Marsh and Merton perpetuate this misconception when they say of these tests that "there are only a few which reject market efficiency." Many of these tests have low power against a fads alternative. The kind of evidence we might expect to find given a fads alternative appears to be there (my 1986 paper, or Lawrence Summers, 1986). There is, it should be noted, also really no reason a priori to

[3]See Kleidon's Table 2, second to the last row. He describes this result as showing that gross violations of the variance inequalities are not improbable. It is true that for $r=.05$, the probability he gives for the log random walk case that $\sigma(p)/\sigma(p^*)>5$ in a sample of 100 is .40. Thus, the main hope for Kleidon's argument would seem to be that the discount rate is lower than was assumed in my paper.

[4]Kleidon has also tested some "conditional" variance inequalities, and did not find evidence of excess volatility.

think that stock prices are excessively volatile relative to the efficient markets model. Just because people probably misprice stocks in alternating waves of interest and disinterest does not necessarily imply that stock prices are more volatile than would be implied by fundamentals. There is some uncertainty about the ultimate significance of the violation of the variance bounds inequalities. But there is no reason to continue putting so little research attention onto behavioral models of financial markets, awaiting the day when evidence against market efficiency is so clear as to allow a jury of randomly chosen citizens to reach a unanimous decision against them. Research should try to be a step or two ahead of that, and this means, I think, working more now to combine modern financial theory with behavioral alternatives.

REFERENCES

Campbell, John Y. and Shiller, Robert J., "Cointegration and Tests of the Present Value Relation," unpublished paper, 1985.

Flavin, Marjorie, "Excess Volatility in the Financial Markets: A Reassessment of the Empirical Evidence," *Journal of Political Economy*, December 1983, *91*, 929–56.

Kleidon, Allan W., "Variance Bounds Tests and Stock Price Valuation Models," *Journal of Political Economy*, forthcoming 1986.

LeRoy, Stephen, and Porter, Richard, "The Present Value Relation: Tests Based on Variance Bounds," *Econometrica*, May 1981, *49*, 555–74.

Lintner, John V., Jr., "Distribution of Incomes of Corporations Among Dividends, Retained Earnings, and Taxes," *American Economic Review Proceedings*, May 1956, *46*, 97–113.

Mankiw, N. Gregory, Romer, David and Shapiro, Matthew D., "An Unbiased Reexamination of Stock Market Volatility," *Journal of Finance*, July 1985, *40*, 677–87.

Marsh, Terry A. and Merton, Robert C., "Dividend Variability and Variance Bounds Tests for the Rationality of Stock Market Prices," *American Economic Review*, June 1986, *76*, 483–98.

Miller, Merton H., "Debt and Taxes," *Journal of Finance*, May 1977, *32*, 261–75.

Myers, Stewart C., "The Capital Structure Puzzle," *Journal of Finance*, July 1984, *39*, 575–92.

Schwert, G. William, "Effects of Model Specification on Tests for Unit Roots," Graduate School of Management Working Paper Series No. GPB 85–10, University of Rochester, 1985.

Shiller, Robert J., (1981a) "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?," *American Economic Review*, June 1981, *71*, 421–36.

______, (1981b) "The Use of Volatility Measures in Assessing Market Efficiency, *Journal of Finance*, May 1981, *36*, 291–304.

______, "Fashions, Fads and Bubbles in Financial Markets," in Jack Coffee and Louis Lowenstein, eds., *Takeovers and Contests for Corporate Control*, Oxford University Press, forthcoming 1986.

Summers, Lawrence, "Do Market Prices Accurately Reflect Fundamental Values?," *Journal of Finance*, July 1986, forthcoming.

West, Kenneth D., "Speculative Bubbles and Stock Price Volatility," Financial Research Center, Memo., No. 54, Princeton University, December 1984.

What Will Take the Con Out of Econometrics? A Reply to McAleer, Pagan, and Volker

By THOMAS F. COOLEY AND STEPHEN F. LEROY*

Our 1981 paper, criticized by Michael McAleer, Adrian Pagan, and Paul Volker (1985), made two points. First, we argued that specification uncertainty renders suspect practically any but the weakest inference about the interest elasticity of money demand. Second, we contended that there is no credible reason to imagine that simultaneity problems are adequately dealt with in existing studies of money demand. Now, if McAleer et al. had wanted to make a truly effective criticism of our paper, they might have pointed out that if the second point is granted, the first does not follow. The Leamer-Leonard method for ascertaining fragility is based on the maintained assumption that the error is orthogonal to all the candidate regressors—precisely the assumption that we criticized in the second half of our paper. Had McAleer et al. argued along these lines, we would have been hard put to come up with a convincing reply. It is true that we suggested (p. 827) that extreme fragility is evidence of serious simultaneity problems. We suspect, however, that this argument would not bear close examination, except perhaps in special cases or as a loose statement of why we were motivated to think about simultaneity problems. Our argument reflected the rhetorical exigencies of a difficult transition, rather than any line of reasoning we could readily make precise. We are surprised that no one has called us on this point.

McAleer et al., of course, could not pursue these lines without invalidating their own purported contribution, which consists of attempting to rehabilitate one-equation-at-a-time estimation despite the fact that the equations being estimated are presumably embedded in simultaneous equation systems. It is therefore no surprise that McAleer et al. ignored our invitation to engage in a serious discussion of macroeconometric practice, given that their ox would be gored more than ours.

McAleer et al. either have an understanding of the nature of simultaneous equations estimation very different from ours, or they completely misunderstood our argument. For example, consider their Section IV, Part A, where they found it strange that we deleted the lagged value of the money stock as an explanatory variable for the current money stock despite our expressed opinion that "such lagged endogenous variables as the lagged money stock...cannot plausibly be excluded from the demand side either explicitly as observable explanatory variables for the demand for money or implicitly from the time dependence of the error" (p. 840). Our intention in the passage just cited, contrary to McAleer et al.'s interpretation, was not to criticize estimates of money demand (such as our own) that exclude the lagged dependent variable. Rather, it was to cast doubt on the presumption that lagged money could plausibly serve as an instrument for the interest elasticity of money demand, as has widely been recommended. Despite the simplicity of this argument, McAleer et al. completely misread it, finding only that it is "strange" that even though we conceded that m_{-1} should in principle enter the money demand equation, we nonetheless suppressed it from an ordinary least squares equation.

I

Let us suppose, contrary to what we argued in our 1981 paper, that simultaneity problems can magically be assumed away. Thus assume that even though we do not know

*University of California, Santa Barbara, CA 93106.
We are indebted to Andrew Rose for helpful comments.

what the correct explanatory variables are, we are nonetheless sure that the unobserved determinant of money demand is statistically independent of these variables. Thus there is no problem with ordinary least squares. These were the conditions assumed in the first half of our paper, and throughout by McAleer et al. Our suggestion was simply that specification uncertainty be explicitly acknowledged, and that the sensitivity of the estimated interest elasticity to respecifications be assessed using the methods developed by Leamer and others. Our idea was to encourage econometricians to report priors explicitly so that readers can compare their own priors to those of the econometrician and evaluate the results accordingly.

This suggestion appeared uncontroversial to us, but apparently not to McAleer et al. They are troubled by the fact that different extreme bounds can result from different classifications of variables as doubtful or free. They conclude that since assessments of fragility depend on a "whimsical" choice of priors, such assessments are altogether unreliable. We are mystified by this criticism. It is indeed true that different priors lead to different posteriors—what would be the point of Bayesian econometrics if it were otherwise? Prior beliefs are by definition treated as given; what is gained by calling them whimsical? McAleer et al. appear to suppose that there is some way to dispense with prior beliefs in doing statistical inference, so that a data set can be made to reveal a single correct inference with the appropriate application of statistical technique. On the contrary, both Bayesian and classical statistics of the Cowles variety depend essentially on prior information. The conclusion that an inference is fragile is not some kind of descriptive fact, as McAleer et al. presume, the validity of which can be impugned if it is shown to depend on a whimsical choice of priors. Rather, the fragility of an inference is conditional on the choice of prior. It is precisely the virtue of Leamer's method, not its fault, that it relates fragility to prior beliefs.

McAleer et al. found fault with extreme bounds analysis because no account is taken of prior beliefs about the signs of coefficients; instead variables are classified only as doubtful or free. Here again we are unpersuaded. One of the most important tasks of empirical econometrics is the verification (or falsification) of sign priors. Now, if the reader knows that in the process of arriving at a preferred specification the econometrician has incorporated his (or her) prior beliefs about the signs of coefficients, as by deleting variables which have "wrong" signs, will he (or she) be persuaded by a report of an estimated equation in which all coefficients have the "right" sign? We doubt it. It is exactly because it is (in many contexts) so easy to find in the parameter space a regression that reproduces prior beliefs that readers are routinely unimpressed by reports of "success" in estimation.

Having demonstrated, at least to our satisfaction, that McAleer et al.'s reservations about extreme bounds analysis are without substance, we must acknowledge that one of their criticisms of our application of extreme bounds analysis is correct, and it is not minor. Just as they suggested, we treated the constant term as doubtful. This was inadvertent; since we are not aware of any theory that could justify suppression of the constant, it should be treated as free. McAleer et al. repeated our calculations with this correction and found that the extreme bounds are near zero. The interpretation is that our data and priors justify a confident conclusion that the interest elasticity of money demand is approximately zero, not that any inference about this parameter is fragile. Since McAleer et al. successfully duplicated our results, we have no doubt that their calculation is correct. But the conclusion that the interest elasticity of money demand is zero reflects the exclusion of lagged terms from our regressions. We suppressed dynamics because their inclusion would only widen extreme bounds which, we (incorrectly) believed, were already very wide. Since we have not recalculated the extreme bounds under a less stringent restriction on priors, we must concede that we have not demonstrated the correctness of our contention that, on reasonable priors, any inference about the interest elasticity of money demand is extremely fragile.

II

The question arises whether McAleer et al.'s rejection of our strictures on conventional estimation practices means that they accept these practices. Although there is no necessary reason for one to imply the other, examination of McAleer et al.'s proposed procedure for arriving at a "tentatively adequate" money demand equation reveals that they do in fact accept received practice, subject to some new bells and whistles which are discussed below. They ignore simultaneity problems, and see nothing undesirable in the informal and unsystematic infusion of prior information during the estimation process. Indeed, they recommend it. Our 1981 paper, of course, criticized received practice on exactly these two points. As an example of McAleer et al.'s casual incorporation of prior information, we need do no more than consider their choice of sample period. Their data set ended twelve years ago. Why not use more recent data? Because "a large number of studies have experienced difficulty in estimating conventional money demand functions for the post-1973 period" (p. 302). This explanation, accompanied by a nod in the direction of financial innovation (the alleged cause of the unruly money demand equations estimated from more recent data), appears to McAleer et al. sufficient to justify truncating the data set to the convenient 1952–73 period for which "satisfactory" estimated money demand equations are a dime a dozen. *Plus ça change, plus c'est la même chose.*

Beginning with this informal choice of an adequate data set, McAleer et al. approach the problem of modeling money demand as a four-step process: 1) begin with an overparameterized model; 2) attempt to restrict the number of parameters in that model using tests for common dynamic structure in the variables; 3) set to zero all coefficients in the restricted model that are insignificant; 4) perform a battery of diagnostic tests to see if the model passes quality control. The first three steps in this process reflect McAleer et al.'s version of David Hendry's (1980, 1983) philosophy of modeling "from the general to the specific." The line of reasoning that underlies this approach is that sequences of properly nested tests can be treated as independent. Thus, by nesting tests in this way, the overall significance level of the tests can be controlled. This feature, however, is not going to be very important when the tests at each stage have low power against reasonable alternatives, as do those used by McAleer et al. Moreover, properly nesting tests requires setting in advance fixed criteria for accepting restrictions, not looking at a collection of t-statistics and making arbitrary decisions about which variables to delete. The end result of the McAleer et al. procedure is a simplified model that represents just one of many possible paths through the thicket of restrictions.

It should also be clear that the general-to-the-specific approach, while it may have some benefits when properly executed, is philosophically at odds with structural econometric estimation. The latter requires one to begin with a model that is subject to identifying and possibly overidentifying restrictions. In the former approach one begins from an overparameterized model and identification and exogeneity restrictions play no particular role. Viewed as a nonstructural model, however, the initial single equation considered by McAleer et al. is too narrowly conceived. Considering that the variables being modeled consist of money, *GNP*, interest rates, inflation rates and wealth, all likely to be jointly endogenous, a vector autoregression would be a more appropriate overparameterized nonstructural model.

The final step of the modeling procedure advocated by McAleer et al. is designed to confront the model with a variety of problems and performance criteria to see how well it holds up. Here the authors make a complete about-face on modeling philosophy. They switch to a specific-to-the-general modeling approach by testing whether the model should be generalized to include a group of variables not considered previously; seasonal variables, additional lags, trends and so on. It is not clear (because it is never discussed) why these variables were not included in the initial overparameterized model. At this stage, McAleer et al. also act as though they have been dealing with a

structural model by reporting the results of a test for simultaneous equation bias even though the test requires candidate instrumental variables, and is consequently applicable only in the presence of identifying restrictions.

Finally, McAleer et al. report the fact that although their model "gave satisfactory performance up to 1973, just like automobiles, age finally caught up with it, and after that date its predictive performance declined dramatically" (p. 305, fn. 11). This might lead the unwary reader to conclude that their equation fails one of the diagnostic tests the importance of which they stress: out-of-sample prediction. On the contrary; McAleer et al. report no problems in this respect. To arrive at such a startling conclusion, they truncated the data set at 1970, reestimated their model, and then compared out-of-sample forecasts with the data through the end of 1973. The fact that the prediction errors had variance comparable to that of the sample errors then led them to report success in the out-of-sample prediction! The breakdown of the model after 1973 apparently has no bearing here. It appears as if McAleer et al. see no need to hold themselves to the demanding standards of model adequacy that they recommend to us.

REFERENCES

Cooley, Thomas F. and LeRoy, Stephen F., "Identification and Estimation of Money Demand," *American Economic Review*, December 1981, *71*, 825–44.

______ **and** ______, "Atheoretical Macroeconometrics: A Critique," *Journal of Monetary Economics*, November 1985, *16*, 283–308.

Hendry, David F., "Econometrics: Alchemy or Science?," *Economica*, November 1980, *47*, 387–406.

______, "Econometric Modelling: The Consumption Function in Retrospect," *Scottish Journal of Political Economy*, 1983, *30*, 193–220.

McAleer, Michael, Pagan, Adrian R. and Volker, Paul A., "What Will Take The Con Out Of Econometrics?," *American Economic Review*, June 1985, *75*, 293–307.

Preference Reversals and the Independence Axiom

By CHARLES A. HOLT*

Of all of the paradoxical patterns of behavior reported in laboratory experiments, one of the most perplexing for decision theorists is the preference reversal phenomenon. A subject in a preference reversal experiment is typically given a choice between a lottery with a high probability of winning a modest amount of money, a "*P* bet," and a lottery with a low probability of winning a large amount of money, a "$ bet." The lowest amount of money for which the subject would willingly sell either of these lotteries is also elicited. The most common "reversal" is for a subject to choose the *P* bet over the $ bet but to put a higher selling price on the $ bet. Preference reversals were first reported by Harold Lindman (1971) and Sarah Lichtenstein and Paul Slovic (1971). In discussing the psychologists' discovery of such reversals, David Grether and Charles Plott remark: "The inconsistency is deeper than mere lack of transitivity or even stochastic transitivity. It suggests that no optimization principles of any sort lie behind even the simplest of human choices..." (1979, p. 623). This paper offers a simple alternative explanation of documented preference reversals, an explanation that does not require the abandonment of the transitivity assumption.

I. Introduction

The dominant theory of individual decision making in stochastic environments is von Neumann-Morgenstern utility theory; its popularity among economists and statisticians is due to its simplicity and its consistency with a wide range of attitudes toward risk. This "expected-utility" theory, however, is not so general that it lacks empirical content, and, indeed, paradoxical behavior that is inconsistent with the theory has been documented for at least thirty years. The most controversial assumption in the von Neumann-Morgenstern theory is the "independence axiom" (sometimes referred to as the substitution axiom). Other axioms involve properties, such as transitivity, completeness, and continuity, that are needed to ensure the existence of a utility functional defined on a space of lotteries, but it is the independence axiom that results in the expected-utility representation that is linear in the probabilities. The independence axiom is an assumption that a lottery X is weakly preferred to a lottery Y, if and only if a compound lottery that yields X with probability q and some other lottery Z with probability $1-q$ is weakly preferred to a compound lottery that yields Y with probability q and Z with probability $1-q$, for *any* lottery Z. Thus the choice between two lotteries, X and Y, is independent of the possible existence of a common (and hence "irrelevant") prospect Z. Human choices between lotteries with hypothetical monetary payoffs have repeatedly, but not overwhelmingly, violated the independence axiom (see, for example, Daniel Kahneman and Amos Tversky, 1979). Also, Raymond Battalio, John Kagel, and Don McDonald (1985) have recently documented a similar pattern of violations of the independence axiom in experiments involving the choices of rats between lotteries involving actual food payoffs.

The experimental results for human subjects have motivated the development of generalizations of von Neumann-Morgenstern utility theory that are based on weaker forms of the independence axiom (for example, see S. H. Chew and Kenneth MacCrimmon, 1979; Peter Fishburn, 1983; and Mark Machina, 1982). The value of this effort to weaken the independence axiom is considerably diminished, however, if more basic

*Department of Economics, Rouss Hall, University of Virginia, Charlottesville, VA 22901. This research was funded by the National Science Foundation under grant SES-82-19720. I am indebted to James Gillespie and Anne Villamil for helpful comments. Peggy Claytor assisted in the preparation of this manuscript.

axioms such as transitivity are violated. The preference reversal phenomenon, in which subjects put a lower selling price on the lottery for which they state a preference, is generally thought to provide such an obstacle. For example, Machina (1982, p. 308) observes that his analysis of expected utility without the independence axiom would have to be further generalized or replaced if such preference reversals are found to be pervasive.

Grether and Plott, in their review of psychologists' preference reversal experiments, concluded that all earlier experiments were irrelevant from an economist's point of view. Most of these experiments did not have monetary or other incentives to motivate subjects to behave rationally, and there was no control for income effects in all of the remaining experiments that did have real incentives. Grether and Plott ran one series of preference reversal experiments without monetary incentives and another series with both monetary incentives and a random lottery-selection device (described below) that controls for income effects. To their evident surprise, they found preference reversals to be even more pervasive in the series *with* monetary incentives. This result is especially noteworthy because it has been shown in other experiments that violations of the Bayes rule in information-processing tasks are common with inexperienced, financially unmotivated subjects, but that such violations are less frequent with experienced, financially motivated subjects (Grether, 1978).

The Grether and Plott results have been replicated by others who have also documented systematic reversals, although the frequency of such reversals has varied with the level of financial motivation and the method of explaining the experimental procedures (Werner Pommerehne, Friedrich Schneider, and Peter Zweifel, 1982; and Robert Reilly, 1982). Among subjects who express a preference for a P bet over the paired \$ bet, the rate of preference reversals in these experiments is typically above 40 percent. To summarize, preference reversals are common, even for financially motivated subjects in carefully conducted laboratory experiments that control for income effects. Grether and Plott note that these reversals constitute an exception to preference theory that "...stands as a challenge to theorists who may attempt to modify the theory to account for this exception without simultaneously making the theory vacuous" (p. 634).

In order to examine the modifications of expected-utility theory that are indicated by observed behavior, it is useful to first clarify the method of controlling for income effects in these experiments. This is done in Section II, where it is shown that income-effect controls used in the experiments are appropriate if the axioms of von Neumann-Morgenstern utility theory are satisfied. Consequently, the observed preference reversals are violations of this expected-utility theory. In Section III it is shown that the method of controlling for income effects used by Grether and Plott is not valid when the independence axiom is violated. Thus observed preference reversals may be due to violations of independence. Section IV contains a comparison of the pattern of preference reversals and the pattern of violations of the independence axiom in some Allais-paradox situations. Section V provides a summary and conclusion.

II. The Random Lottery-Selection Procedure with von Neumann-Morgenstern Utility

The assumption used in this section's analysis of income-effect controls is that the axioms of expected-utility theory hold, so the subject's preferences are represented by the expected value of a von Neumann-Morgenstern utility function denoted $u(\cdot)$. The discussion of income-effect controls pertains to both the Grether-Plott and Reilly papers. (Pommerehne et al. used similar controls, but the exact relationship between actual monetary payments and the lottery outcomes that generated "play money" prizes was not known with certainty by subjects in their experiments.)

In the description of lotteries that follows, the notation $(q, X;\ 1-q, Y)$ will indicate a probability q of obtaining a prize X and a probability $1-q$ of obtaining a prize Y, where the prizes themselves may be lotteries. The experiments reported in both papers involved six matched pairs of lotteries, each

with a P bet and a \$ bet. One pair of bets that was used in both papers, for example, is (11/12, \$2.00; 1/12, −\$2.00) for the P bet and (1/2, \$5.00; 1/2, −\$1.50) for the \$ bet.

The selling prices for lotteries were elicited using the Becker/DeGroot/Marschak method (see Gordon Becker et al., 1964): The subject first reports a selling price for a lottery, and then a random "offer" from a uniform distribution is generated. The subject receives a dollar amount equal to the offer if the offer exceeds the reported selling price, and the subject receives the lottery itself if the offer falls below the reported selling price. It is intuitively apparent that it is in the subject's own interest to report truthfully the lowest amount of money that the subject would willingly accept in exchange for the lottery (the "ask price"): To state a selling price above the subject's true ask price might result in the subject keeping the lottery when it could have been sold for a price above the true ask price, and to state a selling price below the subject's true ask price might result in the subject selling a lottery at a price below the ask price. These are the only effects of reporting a selling price that differs from the ask price, so it is in the subject's own best interest to reveal the ask price. (This result can also be derived from the analysis of the expected-utility-maximizing selling price to be reported.) This revelation property is independent of the subject's attitudes toward risk.

Note that each possible value of the selling price that the subject can report will determine a lottery. Suppose that each elicitation lottery is actually played immediately after the subject chooses a selling price for that lottery, and that the resulting money payment is made before proceeding. Then the subject's wealth at the beginning of the second elicitation procedure is likely to be different from the subject's wealth at the beginning of the first elicitation procedure. To prevent wealth effects from contaminating the data, only one lottery, selected at random, was actually used to determine the subject's payoff at the end of the experiment. More precisely, the subject's choice in each of the six paired P and \$ bets determined a lottery (the one chosen by the subject), and the subject's choice of selling price in each of the twelve elicitation procedures determined a lottery. Thus there were eighteen lotteries determined by a subject's choices, and only one of these lotteries, selected randomly and equiprobably, would be used at the end of the experiment to determine the subject's actual monetary reward. A nonrandom amount of \$7 was added to the outcome to ensure a positive payment.

The analysis of this random lottery-selection method will be for the simplest case of one matched pair of P and \$ lotteries and two elicitation lotteries; the generalization of the argument will be obvious. Let the P bet be denoted by X_p, and let the random payment determined by this bet be denoted $\tilde{x}_P$. The analogous notations for the \$ bet are $X_\$$ and $\tilde{x}_\$$. Thus the P bet is preferred to the \$ bet by an expected-utility-maximizing subject, if and only if $E\{u(w+\tilde{x}_P)\}$ exceeds $E\{u(w+\tilde{x}_\$)\}$, where w denotes initial wealth and $E(\cdot)$ indicates an expectation with respect to the appropriate random variable. The notation for the Becker/DeGroot/Marschak elicitation of a selling price r for a lottery X will be: $B(r; X)$ for the lottery and $\tilde{b}(r; X)$ for the resulting random payment.

The subject's decision problem is to choose the selling prices, $r_\$$ and r_P, and a random variable $\tilde{x} \in \{\tilde{x}_\$, \tilde{x}_P\}$ to maximize:

$$\text{(1)} \quad \frac{1}{3}E\{u(w+\tilde{x})\} + \frac{1}{3}E\{u(w+\tilde{b}(r_\$; X_\$))\} + \frac{1}{3}E\{u(w+\tilde{b}(r_P; X_P))\},$$

because the probability that each of the three lotteries will be relevant at the end of the experiment is 1/3.

It is apparent from the separable form of the expected-utility expression in (1) that the choices of $\tilde{x}$, $r_\$$, and r_P can be made independently. Thus the random variable $\tilde{x}$ selected will be the preferred lottery, and each selling price will actually be the lowest

price at which the subject would willingly sell the lottery. The implication is that the random lottery-selection procedure provides a proper control for income effects if the subject's preferences can be represented by a von Neumann-Morgenstern utility function.

III. The Random Lottery-Selection Procedure without the Independence Axiom

In the previous section's derivation of the objective in equation (1), it was the independence axiom that permitted the objective to be written as an expected value, linear in the lottery-selection probabilities (the "1/3" terms). It was this linearity that resulted in the separability of the three choices. The independence axiom is not assumed to hold in this section.

In the actual experiments, the lottery choices for some pairs were made by the subject after the elicitation of selling prices for those lotteries, and the procedure was reversed for other pairs of lotteries. For this section's analysis of the simple case of only one pair of lotteries, suppose first that the two elicitations precede the lottery choice. Let $\hat{r}_\$$ and $\hat{r}_P$ denote the selling prices that were selected in the elicitation stage. Then the choice between X_P and $X_\$$ in the final state is a choice between the compound lotteries

$$(2)\quad \left[\frac{1}{3},\, X_P;\ \frac{1}{3},\, B(\hat{r}_\$;\, X_\$);\ \frac{1}{3},\, B(\hat{r}_P;\, X_P)\right]$$

and

$$(3)\quad \left[\frac{1}{3},\, X_\$;\ \frac{1}{3},\, B(\hat{r}_\$;\, X_\$);\ \frac{1}{3},\, B(\hat{r}_P;\, X_P)\right]$$

Let the lottery Z be defined:

$$(4)\quad Z \equiv \left[\frac{1}{2},\, B(\hat{r}_\$;\, X_\$);\ \frac{1}{2},\, B(\hat{r}_P;\, X_P)\right].$$

It follows from (2), (3), and (4) that the choice between X_P and $X_\$$ in the final stage is a choice between

$$(5)\quad \left(\frac{1}{3},\, X_P;\ \frac{2}{3},\, Z\right) \text{ and } \left(\frac{1}{3},\, X_\$;\ \frac{2}{3},\, Z\right)$$

The direct implication of the independence axiom is that the left-hand compound lottery in (5) is weakly preferred to the right-hand compound lottery in (5), if and only if X_P is weakly preferred to $X_\$$. The choice between the P bet and the \$ bet in the final stage of the preference reversal experiments is actually a choice between the compound lotteries in (5), and if the independence axiom does not hold, the bet that the subject chooses in the final stage may not be the bet that is actually preferred by the subject, that is, the bet that would be chosen in a direct choice between X_P and $X_\$$. The bet that is actually preferred in such a direct choice may be given a higher selling price in the experiment, even though it is not selected in the pairwise comparison of the compound lotteries. If this occurs, a violation of the independence axiom may result in choices that appear to violate transitivity.[1]

The preceding discussion pertained to the case in which elicitation of the subject's selling prices is done before the subject makes the choice between the P bet and the \$ bet. If, instead, the choice between these two bets is to be made before the elicitation of selling prices, then the choice between X_P and $X_\$$ in the first stage is a choice between the compound lotteries in (5), where Z is a lottery that in this case represents the subject's beliefs about the possible outcomes in the subsequent elicitation stage. As before, a violation of the independence axiom may result in preference reversals in which the bet in the compound lottery selected is the bet with the lowest selling price. Of course, it is possible to specify a utility functional that does not satisfy independence and for which the choices of the lottery and selling prices are independent of the order in which these choices are made. The point being made here is that the lottery chosen (by the individual whose preferences are represented by this utility functional) may not be the lottery that would be preferred in the absence of the elicitation procedures. Another potential

[1] This point has been made independently by Edi Karni and Zvi Safra (1984).

TABLE 1—PAIRED BETS USED IN THE KAHNEMAN AND TVERSKY EXPERIMENTS

Experimental Design	(1)	(2)[a]	(3)	(4)	(5)[a]	(6)
Problems 3; 4	(1.0, £3000)	> (80%)	(.8, £4000; .2, £0)	(.25, £3000; .75, £0)	< (65%)	(.2, £4000; .8, £0)
Problems 3′; 4′	(1.0, −£3000)	< (92%)	(.8, −£4000; .2, £0)	(.25, −£3000; .75, £0)	> (58%)	(.2, −£4000; .8, £0)
Preference Reversal	X_p	<	$X_\$$	$(q, X_p; 1-q, Z)$	>	$(q, X_\$; 1-q, Z)$

[a]The numbers in parentheses indicate the percentage of the 95 subjects' choices that corresponded with the preference directions above the percentage numbers.

source of reversals if the independence axiom does not hold is the possibility that an elicited selling price may not actually be the lowest price that the subject would willingly accept for the lottery because the choice of bets and the choices of selling prices are no longer separable as was the case in equation (1).

IV. A Comparison of Preference Reversals and Allais Paradoxes

A comparison of equations (2) and (3) indicates that the elicitation of selling prices reduces the impact (in a probabilistic sense) of the choice between the P and \$ bets on the outcome of the experiment. This reduction will be called a dilution. If the independence axiom is not satisfied, a dilution, which is due to the random lottery-selection procedure, may cause the reversal in which the subject chooses the compound lottery containing the least preferred bet. In this section, I will argue that this dilution results in the same kind of reversal that is commonly observed in well-known Allais paradoxes.

Kahneman and Tversky report the results of individuals' choices between the pairs of lotteries shown in the first two rows of Table 1. The prizes in these rows were *hypothetical* and were expressed in Israeli pounds. The preference inequality symbols in columns 2 and 5, > and <, indicate the direction of preference for the majority of the 95 subjects that were given these choices. The percentage number in parentheses under each preference inequality indicates the percentage of subjects making the choice indicated by the preference inequality.

Consider the Kahneman-Tversky problems 3 and 4 that are shown in the first row of the table. Most subjects preferred the safe option in column 1 to the risky prospect in column 3. The prospect in column 4 is a compound lottery with a .25 probability of the safe outcome (£3000) in column 1 and a .75 probability of a zero payoff. Similarly, the prospect in column 6 is an analogous dilution of the risky prospect in column 3 with the high probability, .75, of a zero payoff. The modal choice in problems 3 and 4 was to select the safe outcome in the simple choice setting but to select the compound lottery in column 6 that is a dilution of the risky prospect in column 3. This is obviously a direct violation of the independence axiom, although the "violation" is irrelevant to economic theory in the sense that hypothetical payoffs were used.

Next, consider the problems 3′ and 4′ (shown in the second row of Table 1) in which the payoffs in the row above have been reflected around zero. This multiplication of payoffs by minus one changed the direction of the preference inequalities in columns 2 and 5: The risky prospect in column 3 is preferred to the safe outcome in column 1, but the dilution of the safe outcome is preferred to the dilution of the risky prospect in this case. This modal choice pattern is again a violation of independence. Kahneman and Tversky also report other lottery pairs for which a reflection of the

payoffs around zero reverses the directions of preference inequalities.

I believe that the change in modal preference directions (that occurs in the shift from problems 3 and 4 to problems 3′ and 4′) is due to the relative attractivness of the zero payoff used to dilute the prospects in columns 1 and 3; in one case, the payoff used in the dilution is the worst outcome, and in the other case, the payoff used in the dilution is the best. Loosely speaking, an increase in the relative attractiveness of the "common prospect," the zero payoff, seems to make the typical subject more risk averse in the sense that the dilution in column 4 of the safe outcome is preferred to the dilution in column 6 of the risky prospect in column 3. This explanation of the change in modal preference directions is based on the discussions in Machina (1982) and Robert Weber (1982) of the general pattern of violations of the independence axiom in similar situations.[2]

The prospects in columns 1 and 3 in the second row of Table 1 can be associated with the P bet, X_P, and the \$ bet, $X_\$$, as indicated in the third row of the table. For example, the \$ bet involves a small probability of the best outcome, zero, and a large probability of the worst outcome. Similarly, the prospects in columns 4 and 6 are compound lotteries with a probability q of the corresponding bet in columns 1 and 3 and a probability $1-q$ of the common prospect Z, which is zero in this case. If the preference inequalities are as shown in the third row, then the subject would select the compound lottery containing the dilution of the P bet, but the subject actually prefers the \$ bet to the P bet in a direct choice between these two bets. This is precisely the type of behavior that is commonly observed in the preference-reversal experiments: the compound lottery formed from the P bet is selected over the compound lottery formed from the \$ bet, but a higher selling price is given for the \$ bet.

The obvious question at this stage is why the pattern of preference inequalities for the preference-reversal interpretation in the third row of Table 1 should correspond to the pattern in the second row instead of the opposite pattern in the first row. Recall that a major difference between the prospects in the first and second rows is that the common prospect, used in constructing the diluted compound lotteries, was the worst outcome among the first-row prospects and the best outcome among the second-row prospects. For the preference-reversal design given in equations (2) and (3), the common prospect Z in equation (4) is a very good outcome in the following sense. In the Becker/DeGroot/Marschak elicitation of the selling price for a lottery, the subject can report a true minimum selling price and ensure a reward of either the lottery itself or of a randomly generated "offer" that exceeds the reported selling price. It is in this sense that the elicitation lottery is better than the lottery for which the selling price is being elicited. The common prospect Z in the equation involves a fifty-fifty chance of each elicitation lottery and it is in this sense that the common prospect Z is a good outcome relative to $X_\$$ and X_P.

This comparison, however, is admittedly speculative because the structure of common prospect in the preference-reversal context is more complicated than the common prospect used to construct the compound lotteries in

[2] In order to understand Machina's explanation, consider the first two rows of Table 1. In the first row, the prospects in cols. 1 and 3 dominate those in 4 and 6, respectively, in the sense of first-degree stochastic dominance, but these dominance relations are reversed in the second row of Table 1 because the "common prospect" used in the dilution is the best outcome in that row. Machina discusses the effects of the relative attractiveness of the common prospect: "An analysis of the above cited evidence of Kahneman and Tversky, Hagen, and MacCrimmon and Larsson similarly reveals a tendency for individuals to violate the independence axiom by ranking the stochastically dominating pair of prospects 'according to' a utility function which is more risk averse than the one 'used' to rank the stochastically dominated pair" (1982, p. 288). Note that Machina's explanation is not a statement that a typical individual's preferences are represented by a utility function with risk aversion that is increasing in wealth. The violation of the independence axiom results in a situation in which there is no *single* von Neumann-Morgenstern utility function that represents choices between both pairs of prospects. Also see Machina (1983) for a thorough discussion of this pattern of violation of independence.

the Allais-paradox situations in Table 1. In particular, a comparison of the second and third rows of Table 1 shows that the lottery in the second row that is analogous to Z is a certain payoff of £0. This certain prospect stochastically dominates both prospects in columns 1 and 3 of the second row, but the lottery Z in the preference-reversal design may not dominate the X_P and $X_\$$ prospects (in the sense of first-degree stochastic dominance). Although there are similarities in the patterns of behavior in the preference-reversal and Allais-paradox experiments, the incentive structures in these two types of experiments are not identical. Nevertheless, the typical patterns of behavior in both cases can be explained by violations of the independence axiom, and the violations in the preference-reversal experiments should be of special interest to economists because actual monetary incentives have been used in many of these experiments.

V. Summary and Conclusion

A preference reversal occurs when a subject chooses one lottery over another but puts a higher selling price on the lottery not selected. Such behavior is generally thought to violate one or more basic assumptions, for example, transitivity, that underlie all formal preference theories that are commonly used by economists.

The focus in this paper is on the random lottery-selection method of controlling for income effects in preference reversal experiments with financially motivated subjects. If the independence axiom of expected-utility theory is not satisfied, then the lottery choice and the selling price elicitation decisions are no longer separable as would be the case in an expected-utility analysis. In this case an individual might value one lottery more highly than another, but choose the less preferred lottery in an experiment in which the consequences of this choice are diluted by the common prospect associated with the elicitation of selling prices. As a consequence, decisions that appear to violate transitivity can result from direct violations of the independence axiom or of some other axiom such as "reduction of compound lotteries." It is well known that many individuals make choices that are direct violations of the independence axiom in other contexts. Therefore *any* theory of rational choice in such contexts must be derived from a set of axioms that does not include or imply the independence axiom, at least not in its usual "strong" form.

Preference reversals could also be generated by intransitivities, but to abandon transitivity would be a drastic step that would make it difficult to construct a formal choice theory with empirical content. The transitivity assumption is needed for the existence of a utility functional that represents preferences over lotteries; independence is a strong assumption about the functional form of this utility functional (that pertains to the linearity with respect to the probabilities). There has been a considerable amount of recent work on preference theories that involve weaker versions of the independence axiom and that permit more general functional forms (see Machina, 1982, and the references therein). The careful documentation of preference reversals made by *financially motivated* subjects in laboratory experiments conducted by Grether and Plott and others enhances the importance of the generalized expected-utility research.

REFERENCES

Battalio, Raymond C., Kagel, John H. and McDonald, Don N., "Animals Choices Over Uncertain Outcomes," *American Economic Review*, September 1985, *75*, 597–613.

Becker, G. M., DeGroot, Morris H., and Marschak, Jacob, "Measuring Utility by a Single-Response Sequential Method," *Behavioral Science*, July 1964, *9*, 226–32.

Chew, S. H. and MacCrimmon, Kenneth, "Alpha Utility Theory, Lottery, Composition, and the Allais Paradox," Working Paper No. 686, University of British Columbia, 1979.

Fishburn, Peter C., "Transitive Measurable Utility," *Journal of Economic Theory*, December 1983, *31*, 293–317.

Grether, David M., "Recent Psychological

Studies of Behavior Under Uncertainty," *American Economic Review Proceedings*, May 1978, *68*, 70–74.

______ **and Plott, Charles R.,** "Economic Theory of Choice and the Preference Reversal Phenomenon," *American Economic Review*, September 1979, *69*, 623–38.

Kahneman, Daniel and Tversky, Amos, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, March 1979, *47*, 263–91.

Karni, Edi and Safra, Zvi, "Preference Reversal and the Theory of Choice under Risk," Working Papers in Economics No. 142, Johns Hopkins University, September 1984, rev. November 1985.

Lichtenstein, Sarah and Slovic, Paul, "Reversal of Preferences Between Bids and Choices in Gambling Decisions," *Journal of Experimental Psychology*, July 1971, *89*, 46–55.

Lindman, Harold R., "Inconsistent Preferences Among Gambles," *Journal of Experimental Psychology*, May 1971, *89*, 390–97.

Machina, Mark J., "'Expected Utility' Analysis Without the Independence Axiom," *Econometrica*, March 1982, *50*, 227–323.

______, "Generalized Expected Utility Analysis and the Nature of Observed Violations of the Independence Axiom," in B. Stigum and F. Wenstop, eds., *Foundations of Utility and Risk Theory with Applications*, Dordrecht: D. Reidel, 1983.

Pommerehne, Werner W., Schneider, Friedrich and Zweifel, Peter, "Economic Theory of Choice and the Preference Reversal Phenomenon: A Reexamination," *American Economic Review*, June 1982, *72*, 569–74.

Reilly, Robert J., "Preference Reversal: Further Evidence and Some Suggested Modifications in Experimental Design," *American Economic Review*, June 1982, *72*, 576–84.

Weber, Robert J., "The Allais Paradox, Dutch Auctions, and Alpha-Utility Theory," Working Paper, J. L. Kellogg Graduate School of Management, Northwestern University, September 1982.

Shifting Comparative Advantage and Senescent Industry Collapse

By JAMES H. CASSING AND ARYE L. HILLMAN*

As the pattern of international comparative advantage shifts over time, some industries naturally decline. However, the contraction of the senescent industry has frequently been markedly uneven. For example, in the 1950's the United States began rapidly to lose its comparative advantage in shoe production. The shoe industry initially received substantial protection, but gradually the protection eroded and very suddenly, in the mid-1960's, the industry contracted violently. Industry contraction subsequently slowed, protection was lowered but stabilized, and the industry achieved its current much-reduced domestic market share. Similar episodes of relative international competitive decline, protection, gradual erosion of protection and output, and then violent contraction have occurred in various light industries (as, for example, matches and toys), and such a sequence currently appears to be underway in some of the smokestack industries such as steel, shipbuilding, and automobiles.

This tendency for sudden industry collapse to occur in the course of industry decline presents somewhat of a curiosity. One might rather expect more or less continuous decline in output and employment as the industry's capital stock is run down and labor is laid off in response to increased import penetration. Of course, sudden industry collapse might be attributable to production nonconvexities. Because of indivisibilities in the installation of productive capacity, sudden increases in import penetration might take place as the output of foreign plants replaces that of domestic plants which close down. However, when production indivisibilities are very important at the industry level, markets tend not to be competitive. Yet many senescent industries are highly competitive, especially given the presence of foreign competition. Similarly, many small firms producing at minimum average cost may give the impression that a small price decline will drive them all out of business. But, of course, industry long-run marginal cost only appears constant for a given industry output level and the general equilibrium supply curve is nonetheless positively sloped. Yet, in the course of industry decline, sudden and abrupt collapses in output and employment have taken place.

Rather than seeking an explanation for industry collapse in the production technology, we focus in this paper on the government policy response that typically accompanies the senescent industry in decline. Since governments, if they have the political will, can always apply commercial policy to arrest industry decline due to import competition, the contraction of an industry might in itself adversely affect the industry by eroding political support. Intuitively this makes good sense because lower domestic output means that the domestic benefits from protection to the industry—and so benefits to the politician—will be lower while the consumption costs to the constituency will be higher and more obvious. Recently a growing body of literature has indirectly suggested just this point. In particular, empirical studies (see the review of the evidence by Kym Anderson and Robert Baldwin, 1981, and the survey by Baldwin, 1984) find that industry size is an important positive determinant of the supply of protection.

The consequences of such "feedback effects" have, however, received no attention. While much analysis has focused on the

*University of Pittsburgh, PA 15260; and Bar-Ilan University, Ramat Gan, Israel 52100, and University of California, Los Angeles, CA. 90024, respectively. We thank Henry Ursprung and two referees for helpful comments.

political economy of tariff policy formation and its effect on industry activity levels, there has been a neglect of the counterdirection as changing industry activity levels themselves reshape political support.[1] We shall incorporate such political-influence feedback considerations into a simple general equilibrium model of adjustment due to sharpened import competition, and show that sudden industry collapse is one of the possible adjustment paths. The possibility of collapse is particularly interesting because the model admits of no nonconvexities in production or consumption, and no discontinuities such as discrete tariff changes in the government's commercial policy actions. Nonetheless, very sharp jumps in economic activity—mathematically, "catastrophes"[2]—emerge due to otherwise smoothly adjusting variables interacting in the adjustment process.

We shall begin in Section I with a portrayal of factor-market adjustment and industry decline as the response to a fall in the world price of an industry's output, with the authorities remaining passive. Policy response is introduced in Section II by means of a positive political-equilibrium relationship between tariff levels and industry activity levels as measured by the size of the labor force. The political supply of protection is expressed in its reduced functional form and accepted as conforming to the theoretical conclusions and empirical findings of Anthony Downs (1957), Mancur Olson (1965), Jonathan Pincus (1975), Baldwin (1976, 1984), and others. Section III then incorporates the political-equilibrium tariff response into industry adjustment and shows how, as a consequence of the political response to industry activity levels, sudden industry collapse can occur even if output prices vary only in a continuous way. We keep the model as neoclassical as possible, and, in particular, convex in production by adopting the familiar two-sector model with capital industry-specific in the short run.

I. Industry Decline in the Absence of Policy Response

The model characterizes dynamic equilibrium conditions in factor markets and in the political market. To describe factor-market equilibrium, we posit a dynamic version of the competitive two-sector specific-factors model for a small country. Labor L_i and capital K_i are employed in industry i to produce output x_i via the constant-returns industry neoclassical technologies[3]

$$(1) \qquad x_i = F^i(L_i, K_i) \qquad i = 1,2.$$

We represent by industry 1 the senescent industry and by 2 the rest of the economy. Aggregate factor supplies, L and K, are fixed. In long-run competitive equilibrium the value of the marginal product of a factor is equated across industries, and hence, with P denoting the relative domestic price of the senescent industry's output,

$$(2) \quad PF_L^1(L_1, K_1) = F_L^2(L - L_1, K - K_1)$$

$$(3) \quad PF_K^1(L_1, K_1) = F_K^2(L - L_1, K - K_1).$$

Since we are interested in the adjustment process, we need to explain how factor reallocation occurs out of long-run equilibrium. We assume therefore that factors are always paid the value of their marginal products and that inputs move to higher reward industries. Labor is mobile between industries, and hence in particular when the world price of the senescent industry's output declines to reduce the value of labor's marginal product in that industry, labor adjusts to find alternative employment elsewhere in the economy (in industry 2). However, capital, which may be regarded as both

[1] See the survey by Baldwin (1984). In Wolfgang Mayer's 1984 analysis of voting equilibria, levels of the tariff affect the demand for tariff increases, and hence implicitly equilibrium tariff levels are related to output levels.

[2] See T. Poston and I. Steward (1978).

[3] Nonincreasing returns to scale would suffice, although this would require the further specification of a fictitious factor for factor payments to exhaust the value of output. In (1), capital and labor are technological complements, and hence $F_{KL}^i > 0$, $i = 1,2$.

human and physical, lacks the same occupational mobility and adjusts much more slowly to a fall in the value of its marginal product.[4] Adjustment of capital is described by the equation of motion

$$\dot{K}_1 = g(r_1 - r_2), \tag{4}$$

where $r_i = P_i F_K^i$, $g(0) = 0$; $g'(\cdot) > 0$.

The value of the domestic relative price P is determined jointly by an exogenous relative world price and intervention by the authorities. Let intervention take the form of a specific tariff.[5] Given the world price P^* and the tariff T, and therefore the domestic price $P = (P^* + T)$, long-run equilibrium employment combinations of factor inputs satisfy

$$\dot{K}_i = \dot{L}_i = 0, \qquad i = 1,2. \tag{5}$$

In (L_1, K_1) space of employment in the senescent industry, equilibrium is accordingly determined by the intersection of two loci with respective positive slopes:

$$\left.\frac{dK_1}{dL_1}\right|_{\dot{L}_i = 0} = -\left[\frac{PF_{LL}^1 + F_{LL}^2}{PF_{KL}^1 + F_{KL}^2}\right] > 0; \tag{6}$$

$$\left.\frac{dK_1}{dL_1}\right|_{\dot{K}_1 = 0} = -\left[\frac{PF_{KL}^1 + F_{KL}^2}{PF_{KK}^1 + F_{KK}^2}\right] > 0. \tag{7}$$

[4]The point is that there are asymmetries in adjustment as some factors exit the declining industry more swiftly than others. For a model which portrays adjustment with one factor mobile and another imperfectly so, see Gene Grossman (1983).

[5]We adopt this assumption for expositional purposes. In our competitive setting, there is tariff/quota symmetry (with respect to either a specific or ad valorem tariff). On the other hand, if the domestic market structure were not competitive, one instrument may dominate the other as the politically preferred means of protection (see our 1985 paper). Also, intervention could entail the use of tax-subsidy instruments (see Michael Mussa, 1982). Jagdish Bhagwati (1982) extends the policy options in the face of shifting comparative advantage to include labor immigration.

Figure 1 illustrates the initial dynamically stable factor-allocation equilibrium (L_1^a, K_1^a) for the relative price of senescent industry output P'.

When the world price P^* declines to reflect shifting comparative advantage, and the tariff is left unchanged so that industry 1's domestic price falls, the factor-allocation loci in Figure 1 shift to reflect reduced factor demand. At $P'' < P'$, the long-run equilibrium employment levels are (L_1^c, K_1^c). Adjustment to the fall in the price of the industry's output begins with labor exiting the declining industry 1, and labor employment falling from L_1^a and L_1^b at the maintained capital stock K_1^a.

It should be observed that the initial employment contraction from L_1^a to L_1^b is not our characterization of collapse, and that this contraction would be only as abrupt as the initial price decline. In order to focus on long-run adjustment to the new external price, we assume that the initial external price decrease is once and for all, rather than introducing another exogenous dynamic equation of motion for the world price P^*.

Once on the equilibrium locus of the faster-adjusting variable (labor), adjustment now proceeds over time according to equation (4) as capital exits the now lower-return declining industry. As capital leaves industry 1, additional labor is reallocated in order to sustain equation (2). Labor reallocation ceases when the industry's capital stock has declined to K_1^c, and conditions (2) and (3) are reestablished.

This process of smooth adjustment assumes, however, that the authorities formulating commercial policy remain passive by not adjusting the tariff when the world price and the output of industry 1 decline. Naturally, a sufficiently large increase in the tariff could altogether eliminate the market-induced incentives for labor to exit from and capital to be run down in industry 1. But which commercial policy response will emerge? Between "none at all" and "complete insulation from the initial stock" depends upon how the supply of protection responds to changes in the senescent industry's activity level.

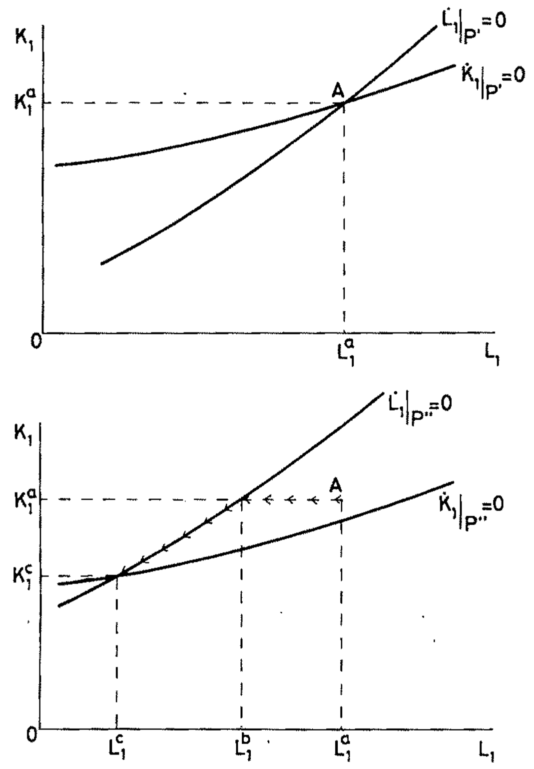


FIGURE 1

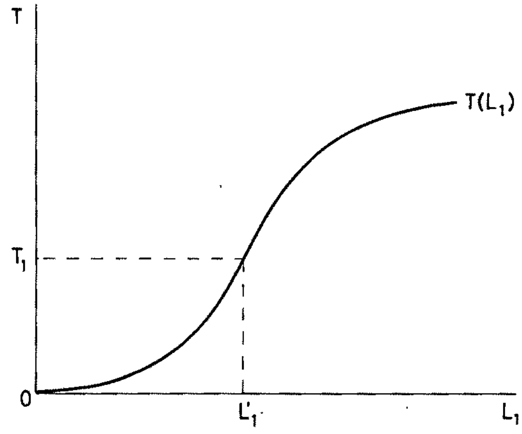


FIGURE 2

II. Political Support and Industry Size

The "supply of protection" depends, among other things, upon the absolute size of an industry. The larger output or employment, the higher the level of protection received by an import-competing industry, other things equal. In order to capture this relationship, we characterize the tariff level as a function of labor employment in industry 1,

$$T = T(L_1), \tag{8}$$

where $T_L > 0$. That is, the height of the tariff increases with the number of jobs provided by the protected industry.[6]

Empirically, the positive relationship between tariffs and size of the industry is well established.[7] The theoretical underpinnings for the qualitative functional form derive from the theory of collective-action protectionist demands and political-support responses. Although political effectiveness increases with size, at the margin a political coalition benefits from being neither too small nor too large. For if small, then while the coalition's protectionist demands may be cohesively articulated, the limited political returns inhibit the authorities from being too forthcoming in increasing the protection already provided. And if overly large, the coalition confronts free-rider problems that hinder cohesive political action. Also the consumption costs of marginal tariff increases become larger and more obvious. Coalitions that are neither too small nor too large are, on the other hand, politically attractive for the dispensing of the benefits of intervention, while at the same time not disadvantaged by free-rider disincentives which limit the effectiveness of political organization (see Olson).

[6]While we emphasize the relation between industry activity levels and protection, other industry characteristics influence industry protection levels. See, for example, the variables considered by Edward Ray (1981), Howard Marvel and Ray (1983), Réal Lavergne (1983), and Paul Godek (1985).

[7]Anderson and Baldwin find that this holds in a general manner across industries for a number of industrialized countries.

These implications of coalition size when brought together imply the industry political-support function characterized by a sigmoid relationship as in Figure 2. At L_1' where there is a point of inflection, the marginal tariff response provided is maximal. For L_1 smaller than L_1', $T_{LL} > 0$; and for L_1 larger than L_1', $T_{LL} < 0$. Equilibrium in the political market is then characterized in the reduced form as the level of tariff protection $T = T(L_1)$ as, for example, T_1 at L_1'.

III. Political-Feedback Effects and the Possibility of Industry Collapse

Now consider the influence of the authorities' policy response on industry adjustment. Substituting (8) into the factor-market equilibrium conditions, we have the analogues of equations (2) and (3) now encompassing the political-market equilibrium:

$$(9)\quad [P^* + T(L_1)]\,F_L^1(L_1, K_1) = F_L^2(L - L_1, K - K_1);$$

$$(10)\quad [P^* + T(L_1)]\,F_K^1(L_1, K_1) = F_K^2(L - L_1, K - K_1).$$

Consequently, with tariff policy endogenized, factor-input combinations which maintain constant labor and capital employment have the respective properties,

$$(11)\quad \left.\frac{dK^1}{dL^1}\right|_{\dot{L}=0} = -\left[\frac{(P^*+T)F_{LL}^1 + F_L^1 T_L + F_{LL}^2}{(P^*+T)F_{KL}^1 + F_{KL}^2}\right] \gtreqless 0;$$

$$(12)\quad \left.\frac{dK^1}{dL^1}\right|_{\dot{K}=0} = -\left[\frac{(P^*+T)F_{KL}^1 + F_K^1 T_L + F_{KL}^2}{(P^*+T)F_{KK}^1 + F_{KK}^2}\right] > 0.$$

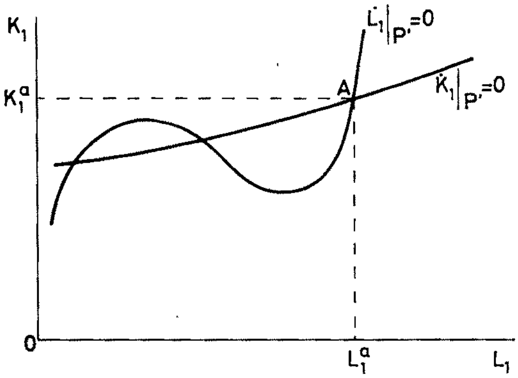


FIGURE 3

The capital-allocation locus thus retains the unambiguously positive slope established in the absence of policy response. But not so the labor-allocation locus, that now contains an intermediate portion that may be negatively sloped. This is because the term $F_L^1 T_L$ is positive and, in particular, reflecting the character of political-market equilibria, the component T_L ranges from lower values for low levels of L_1 to higher values for intermediate levels of L_1 and finally back to lower levels for high levels of L_1.[8] But when the positive term dominates the numerator of equation (11), the labor-equilibrium locus will have a negatively sloped portion. The (stable) equilibrium in (L_1, K_1) space is then as depicted in Figure 3, where both market-equilibrium loci incorporate changing tariff policy.

Consider now an initial equilibrium at the point A in Figure 3, and suppose that the world price P^* were to fall. We again take the world price decline to be discrete, although a continuously falling world price

[8]The critical term $F_L^1 T_L$ depends upon the technology and the reduced-form tariff formation function. Define $\alpha(L_1) = F_L^1 T_L$. Then $\alpha > 0$ and reaches its maximum at that L_1 where $\alpha' \equiv T_L F_{LL}^1 + F_L^1 T_{LL} = 0$, while F_L^1 falls unambiguously as L_1 increases; the term T_L at first increases with L_1 and can legitimately range arbitrarily high at the inflection point L_1' in Figure 2. However, beyond L_1' both T_L and F_L^1 fall as L_1 increases. Thus, the range of interest in Figure 3 corresponds to the range in Figure 2 in the vicinity of the inflection point of the sigmoid function.

would generate the same shifted loci eventually. Unless the authorities raise the tariff sufficiently to offset the world price decline, the domestic price will fall as well, and the equilibrium loci will shift to reflect the reduced values of the marginal products of both factors.

When a political-support maximizing tariff is set so as to trade off the interests of gainers and losers from protection, the prediction is that a tariff-setting authority will permit the domestic price to decline at least partially in response to a fall in the world price.[9] This accords with the characteristic of political equilibrium that the authorities' policy response ensures that at least some benefit from the fall in the world price is passed on to those who have lost as a consequence of industry protection. Accordingly, with full tariff offset for the decrease in world price not provided, at least some benefit from the price decline is dispersed throughout the economy to maintain a marginal balancing of political support, and the industry begins its decline to a new equilibrium. Figure 4 illustrates the path of industry adjustment. Initially, beginning from A, the exit of labor reduces industry employment to L_1^b. Eventually, capital also begins to be run down in the industry and, as the capital stock decreases, more labor exits so that labor employment declines smoothly along the labor market equilibrium locus. However, when the capital stock has been reduced to K_1^c, the relatively faster adjusting variable, labor, reaches a fold at L_1^c. A further small decline in the capital stock below K_1^c thus induces an abrupt collapse in industry employment of labor, from L_1^c to L_1^d.[10]

This collapse comes about as a consequence of the accelerating decline in political effectiveness and hence loss of support in the political market for protection. As adjust-

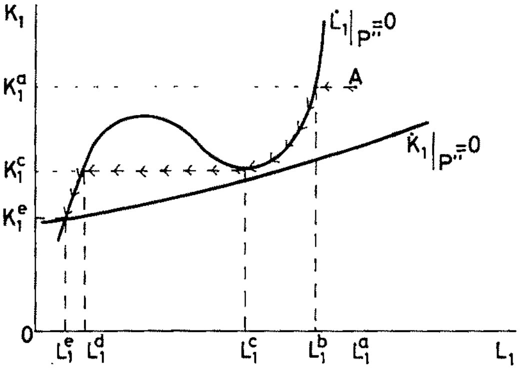


FIGURE 4

ment continues and employment contracts, fewer individuals remain to have their jobs protected by government intervention. The industry's political weight is diminished and the tariff falls. At L_1^c, protection decreases so rapidly that a further smooth reduction in employment of labor could only maintain equality of the value of labor's marginal product across industries if the capital stock were to increase. This, of course, will not happen in the declining industry. So the abrupt collapse of employment to L_1^d takes place. Subsequently, beyond L_1^d, the industry returns to a path of gradual smooth contraction until equilibrium consistent with the fall in world price and shift in comparative advantage is reached at (L_1^e, K_1^e).

The sudden exodus of the most footloose inputs—new entrants, low-seniority workers, and generally less-industry-specific inputs —will have a dual image in the rapid decline in industry output, and in the rate of return to slower-adjusting inputs. This is obviously at least casually consistent with the plight of the senescent industry and its region. The younger workers leave; the more skilled older workers are suddenly asked to take pay cuts and see their wages falling; the physical equipment and plants have their capital values suddenly written down. Ultimately, though, the contraction ends and the industry output stabilizes at a new, albeit lower, level.

If independently of size industries move along the labor-equilibrium locus at the same speed, a larger industry requires a longer

[9]See Sam Peltzman (1976) and, more specifically with reference to the optimizing political-influence protectionist response to senescent-industry decline, Hillman (1982).

[10]This is the "catastrophe" of catastrophe theory.

period of time to reach the stage of rapid decline than a smaller industry. This is also consistent with the stylized facts. For example, a smaller industry such as footwear receives less protection initially and is in or closer to the area where political support drops off quickly as it declines in response to changing comparative advantage. It therefore reaches the stage of abrupt decline sooner than a larger industry such as textiles.[11]

IV. Concluding Remarks

Senescent industries do appear to have exhibited the phenomenon of output and employment collapse in competitive environments where nonconvexities due to production indivisibilities appear not to have been important. The analysis of feedbacks in political-support protectionist responses in the course of industry decline presents an explanation of such collapse. The dependence of the effectiveness of lobbying activity on coalition size yields a political-equilibrium relation between industry size and the tariff level which politically wary authorities will provide. Incorporating this political response into the adjustment process has revealed the possibility of industry collapse based on the character of political-market equilibria derived from the theory of relative effectiveness of diversely sized political coalitions.

It is probably not surprising that senescent industries decline, because in the course of their decline they lose their political support, so feeding back to exacerbate their decline. For example, any efforts made appear half-hearted and no one really tries to protect the shoe industry in New England any more. The surprise is the *sudden* collapse. Here we provide one plausible explanation for such a phenomenon by highlighting the potentially important but neglected linkages between changing industry characteristics and shifting political responsiveness during an adjustment episode.

[11]We thank a referee for this observation.

REFERENCES

Anderson, Kym and Baldwin, Robert E., "The Political Market for Protection in Industrial Countries," Staff Working Paper No. 492, World Bank, 1981.

Baldwin, Robert E., "The Political Economy of Postwar U.S. Trade Policy," *The Bulletin*, Graduate School of Business Administration, New York University, 1976.

______, "Trade Policies in Developed Countries," in Ronald Jones and Peter Kenen, eds., *Handbook of International Economics*, Vol. 1, Amsterdam: North-Holland, 1984, 571–619.

Bhagwati, Jagdish N., "Shifting Comparative Advantage, Protectionist Demands, and Policy Response," in his *Import Competition and Response*, Chicago: University of Chicago Press, 1982, 153–84.

Cassing, James H. and Hillman, Arye L., "Political Influence Motives and the Choice between Tariffs and Quotas," *Journal of International Economics*, November 1985, *19*, 279–90.

Downs, Anthony, *An Economy Theory of Democracy*, New York: Harper and Row, 1957.

Godek, Paul E., "Industry Structure and Redistribution through Trade Restrictions," *Journal of Law and Economics*, October 1985, *28*, 687–703.

Grossman, Gene, "Partially Mobile Capital: A General Approach to Two-Sector Trade Theory," *Journal of International Economics*, August 1983, *15*, 1–17.

Hillman, Arye L., "Declining Industries and Political-Support Protectionist Motives," *American Economic Review*, December 1982, *72*, 1180–87.

Lavergne, Réal P., *The Political Economy of U.S. Tariffs: An Empirical Analysis*, New York: Academic Press, 1983.

Marvel, Howard P. and Ray, Edward J., "The Kennedy Round: Evidence on the Regulation of International Trade in the United States," *American Economic Review*, March 1983, *73*, 190–97.

Mayer, Wolfgang, "Endogenous Tariff Formation," *American Economic Review*, December 1984, *74*, 970–85.

Mussa, Michael, "Government Policy and the Adjustment Process," in Jagdish N. Bhagwati, ed., *Import Competition and Response*, Chicago: University of Chicago Press, 1982, 73–120.

Olson, Mancur, *The Logic of Collective Action: Public Goods and the Theory of Groups*, Cambridge: Harvard University Press, 1965.

Peltzman, Sam, "Toward a More General Theory of Regulation," *Journal of Law and Economics*, August 1976, *18*, 211–40.

Pincus, Jonathan J., "Pressure Groups and the Pattern of Tariffs," *Journal of Political Economy*, August 1975, *83*, 757–78.

Poston, T. and Steward, I., *Catastrophe Theory and Its Applications*, London: Pittman, 1978.

Ray, Edward J., "The Determinants of Tariff and Nontariff Trade Restrictions in the United States," *Journal of Political Economy*, February 1981, *89*, 105–21.

Measures of Risk Aversion and Comparative Statics of Industry Equilibrium

By ELIE APPELBAUM AND ELIAKIM KATZ*

In recent years there has been a large volume of papers regarding the firm under uncertainty. These papers extended the works of Agnar Sandmo (1971) and Hayne Leland (1972) in various directions. For example, Raveendra Batra and Aman Ullah (1974) looked at the inputs decision of the firm under uncertainty, Katz, Jacob Paroush, and Nava Kahana (1983) considered the price discriminating firm under uncertainty, Katz (1984) examined the optimal location of the firm under uncertainty, and so on.

One of the major contributions of Sandmo's work was the finding that comparative statics results for the firm under uncertainty depend on absolute risk aversion and/or relative risk aversion. Thus, Sandmo found that the effects of changes in fixed costs, mean price, and price variability depend on whether absolute risk aversion is increasing or not, and that the effect of a change in a profit tax depends on whether relative risk aversion is decreasing or not.

The dependence of responses of firms under uncertainty on absolute and/or relative risk aversion has been repeatedly confirmed in the literature for different scenarios. In general, it was shown that assumptions about measures of risk aversion are both necessary and sufficient for the derivation of conclusive comparative statics results.

It is the purpose of this paper to examine the extent to which such dependence is still true in a model in which the number of firms in the market is determined endogenously. In other words, we attempt to find out the extent to which absolute and/or relative risk aversion influence matters in industry equilibrium. It is shown that once a competitive equilibrium is introduced, the dependence of comparative statics on absolute/relative risk aversion all but disappears. Assumptions about absolute and/or relative risk aversion are either unnecessary or insufficient to sign the effects of changes in fixed costs, mean price, price variability and the tax rate on the output of the individual firm, on industry output, and on the number of firms in the industry. On the other hand, the effect of a change in the reservation utility on the equilibrium size of the firm *is* shown to depend on absolute risk aversion. We also compare market equilibrium with and without uncertainty, and show that whereas aggregate output is smaller with uncertainty, individual firms' output and the number of firms may be either larger or smaller.

I

Consider a competitive industry of n identical firms, each of which produces a homogeneous output q_i. The industry demand is stochastic and given by

$$p = f(Q, \lambda) + \gamma\varepsilon, \tag{1}$$

where Q is total industry output, p is the price, $\partial f/\partial Q < 0$, ε is a random variable such that $E(\varepsilon) = 0$, $E(\varepsilon^2) = 1$, and λ and γ are shift parameters such that $\partial f/\partial\lambda > 0$ and $\gamma > 0$. Given the symmetry of the firms, we have

$$Q = \sum_{i=1}^{n} q_i = nq. \tag{2}$$

Let each firm's profits be given by

$$\pi(q) = pq - C(q) - k, \tag{3}$$

where k is fixed costs associated with any

*York University; Bar Ilan University and York University, Downsview, Ontario M35 1P3 Canada. We thank John Riley and two referees for useful comments and suggestions.

strictly positive output level and $C(q)$ is the variable cost function which is assumed to be increasing, with $C(0)=0$.

Firms are assumed to be maximizing the expected utility of profits $E[U(\pi)]$ where $U(\cdot)$ is a von Neumann-Morgenstern utility function such that $U'>0$ and $U''<0$ (implying risk aversion). Hence, each firm sets its output by solving

$$(4) \qquad \operatorname*{Max}_{q} E[U[pq-C(q)-k]],$$

yielding the first-order condition

$$(5) \qquad E[U'(\pi)(p-C'(q))]=0,$$

and also yielding the second-order condition that $\partial^2 E[U(\pi)]/\partial q^2<0$.

In addition, the number of firms in the industry is determined by free entry and exit, such that in equilibrium the expected utility of being in the industry is equal to the utility of some benchmark activity. Denoting the utility of this benchmark activity by b yields the industry equilibrium condition

$$(6) \qquad E[U[pq-C(q)-k]]-b=0.$$

The equilibrium of our model is, therefore, determined by the intersection of equations (5) and (6).

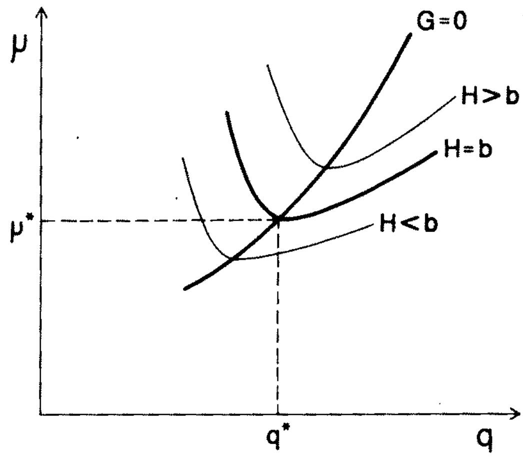
It is useful to characterize the industry equilibrium in the (q,μ) space, where $\mu \equiv E(p)=f(Q)$. To do this, we first define the locus of (q,μ) combinations which satisfy the firms' optimality condition as

$$(7) \qquad G(q,\mu)\equiv E[U'[(\mu+\gamma\varepsilon)q-C(q)-k] \cdot[\mu+\gamma\varepsilon-C'(q)]]=0.$$

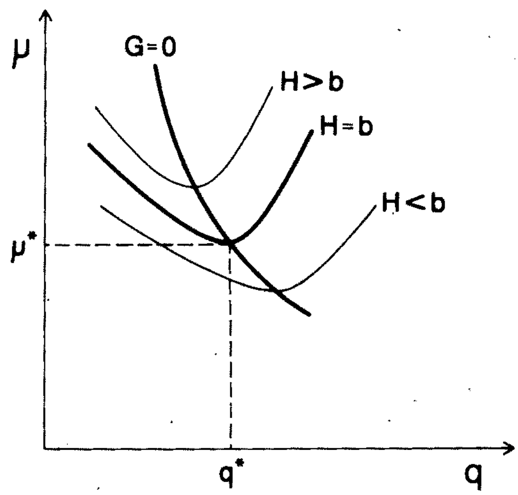
The slope of this curve is given by

$$(8) \qquad d\mu/dq|_{G=0}=-(G_q/G_\mu) = -G_q/(E[U''(\cdot)q(\mu+\gamma\varepsilon-C'(q)] + E[U'(\cdot)]).$$

Since $G_q=E[U''(\cdot)[\mu+\gamma\varepsilon-C'(q)]^2]-C''(q)E[U'(\cdot)]=\partial^2E[U(\pi)]/\partial q^2<0$ (from



(a)



(b)

FIGURE 1

the firms' second-order condition),[1] we know that the sign of the slope of $G=0$ is the same as the sign of G_μ.

It is well known from the work of Sandmo that, if absolute risk aversion is decreasing,

[1]It should be noted that, given risk aversion, the convexity of $C(q)$ is a sufficient but not a necessary condition for $G_q<0$. Thus, instead of assuming convexity of C, we only need to make the weaker assumption that the second-order condition for (4) holds.

then $E[U''(\cdot)(\mu+\gamma\varepsilon-C'(q)]$ is positive. In such a case the $G=0$ curve has a positive slope as illustrated in Figure 1a. If, however, absolute risk aversion is increasing, then the denominator in (8) may (but need not) be negative, as illustrated in Figure 1b.

Next we define the firms' iso-expected utility curves by

$$(9)\quad H(q,\mu)\equiv E\big[U[(\mu+\gamma\varepsilon)q - C(q)-k]\big]=\overline{U}.$$

For each level of utility $H(q,\mu)=\overline{U}$ defines the locus of (q,μ) combinations which yield the utility level $\overline{U}$. To find the slope of these curves, we define $q(\mu)$ as the value of q which (for a given μ) solves $G(q,\mu)=0$, that is, the optimal output when expected price is μ. Then, since $G(q(\mu),\mu)=0$ and $G_q<0$ (from the second-order condition), we must have

$$(10)\quad H_q=G(q,\mu)\lesseqgtr 0 \quad \text{if} \quad q\gtreqless q(\mu).$$

Thus, since $H_\mu=qE[U'(\cdot)]>0$, the slope of the iso-expected utility curves is given by

$$(11)\quad \left.\frac{d\mu}{dq}\right|_{H=\overline{U}}=-\frac{G(q,\mu)}{H_\mu}\gtreqless 0 \text{ if } q\gtreqless q(\mu)$$

so that the iso-expected utility curves are U-shaped as shown in Figure 1. Clearly, higher curves represent a higher level of expected utility.

In the short run, for any given μ, the firms will choose the corresponding output on the $G=0$ curve (i.e., $q(\mu)$). The corresponding level of expected utility is given by the iso-expected utility curve going through the point $(q(\mu),\mu)$ on the $G=0$ curve. In Figure 1, for any $\mu<\mu^*$ the firms' expected utility is less than the reservation utility b, whereas for any $\mu>\mu^*$, the firm's expected utility is greater than b (along the $G=0$ curve).

Industry equilibrium is defined by the pair (q^*,μ^*) which lies on the $G=0$ and $H=b$ curves. That is, in industry equilibrium the firms' expected utility will be equal to the reservation utility b. This will be brought about by free entry and exit of firms with corresponding changes in industry output and hence expected price.

To examine the stability of the system, we make the usual assumptions about the response of firms and the industry when out of equilibrium. We assume that when H exceeds b, new firms enter the industry and reduce the mean price and when H is smaller than b firms leave the industry and cause the mean price to rise. Thus $d\mu/dt\gtreqless 0$ if $H\lesseqgtr b$. Furthermore, we assume that when the marginal expected utility of profits is negative, firms reduce their output and when the marginal expected utility is positive, firms increase their output. Thus $dq/dt\gtreqless 0$ if $G\gtreqless 0$.

Writing the dynamics of the system as

$$(12a)\quad dq/dt=\lambda_1 G(q,\mu);$$

$$(12b)\quad d\mu/dt=-\lambda_2[H(q,\mu)-b],$$

where λ_1,λ_2 are positive adjustment parameters, which without loss of generality can be set equal to unity, we have that the system is locally stable if trace $[J]<0$ and $|J|>0$ where

$$(13)\quad J\equiv\begin{bmatrix} G_q & G_\mu \\ -H_q & -H_\mu\end{bmatrix}=\begin{bmatrix} G_q & G_\mu \\ 0 & -H_\mu\end{bmatrix}.$$

Hence, since trace $[J]=G_q-H_\mu<0$ and since $|J|=-G_qH_\mu>0$, we have a stable system. Indeed, since $[\text{trace }[J]]^2-4|J|>0$, we have a stable node.

Finally, it is interesting to compare the equilibrium with and without uncertainty (and with the case of risk-neutral firms). Assuming that under certainty firms face the expected price (with certainty) and taking $b=0$ with $U(0)=0$, we get that the $G=0$ is then simply the marginal cost curve, whereas the $H=0$ curve is then the average cost curve. But, since from the concavity of U we have that $E\{U[\pi(q)]\}<U[E(\pi)]$, we know that the average cost curve is lower than the $H=0$ curve. Moreover, since from (4) we have $\mu=C'(q)-\gamma\,\mathrm{Cov}[U'(\cdot),\varepsilon]/E[U'(\cdot)]>C'(q)$, the marginal cost curve is also lower than the $G=0$. Thus, we conclude that the expected price under certainty (or risk neu-

trality) will be lower than under uncertainty, and hence total industry output must be higher under certainty. Surprisingly, however, the output of individual firms and the number of firms may be either higher or lower under uncertainty. In industry equilibrium under uncertainty it is possible, therefore, to have either "excess capacity," or "undercapacity."

II

In this section we consider some comparative statics results of our model. A change in any parameter will shift both the $H=b$ and $G=0$ curves. Hence, to determine the effects of a change in a given parameter we have to know the direction and the relative shift of the curves. Thus, for example, consider the case depicted in Figure 1a. A change in a parameter will cause q^* to increase, if and only if, at the original output level q^*, the $H=b$ curve shifts up more (down less) than the $G=0$. In the case depicted in Figure 1b the opposite is true. Once these shifts are known, the comparative statics results follow immediately.

PROPOSITION 1: *An increase in fixed costs reduces industry output and the number of firms in the industry, but raises the output of firms remaining in the industry. This result is independent of absolute and/or relative risk aversion.*

PROOF:

Denoting the initial equilibrium output by q^*, we have from (7) and (9)

$$\left.\frac{d\mu}{dk}\right|_{\substack{H=b\\q=q^*}} = -\frac{H_k}{H_\mu} = \frac{E[U'(\cdot)]}{q^*E[U'(\ )]} = \frac{1}{q^*} \tag{14}$$

and

$$\left.\frac{d\mu}{dk}\right|_{\substack{G=0\\q=q^*}} = -\frac{G_k}{G_\mu} \tag{15}$$

$$= \frac{E[U''(\cdot)(p-C')]}{q^*E[U''(\ )(p-C')]+E[U'(\ )]}$$

and hence

$$A \equiv \left.\frac{d\mu}{dk}\right|_{\substack{H=b\\q=q^*}} - \left.\frac{d\mu}{dk}\right|_{\substack{G=0\\q=q^*}} \tag{16}$$

$$= \frac{E[U'(\ )]}{q^*[E[U''(\ )(p-C')]q^*+E[U'(\ )]]}.$$

Thus, if the denominator of A is positive then an increase in k shifts the $H=b$ curve upwards more than the $G=0$ curve. This case corresponds to Figure 1a, and as can be seen from that figure (this is left to the reader), the increase in k clearly raises q and μ.

Alternatively, if the denominator of A is negative, the $H=b$ curve shifts upwards by less than the $G=0$ curve. This case corresponds to Figure 1b. Hence, as is seen from that figure, once again the increase in k raises q and μ. The increase in μ implies, of course, a smaller number of firms and a lower industry output.

This result should be contrasted with Sandmo's result that the individual firm's output will fall as fixed costs rise, providing absolute risk aversion is decreasing. Within the context of our diagram, the Sandmo result can be obtained by noting that in the absence of an industry equilibrium condition, μ is given. Hence only the $G=0$ curve shifts in accordance with equation (15) yielding Sandmo's result.

The rationale for our result is found in equation (6). This equation requires that in equilibrium a firm remaining in the industry should be no better or worse off than it was before the change in fixed costs. Thus, in equilibrium its degree of absolute risk aversion will remain unchanged and hence, the way in which risk aversion would alter in response to a change in the firm's welfare does not matter.

PROPOSITION 2: *A spread-preserving increase in demand raises industry output and the number of firms in the industry, but leaves unchanged the output of individual firms. This*

result is independent of absolute and/or relative risk aversion.

PROOF:

The demand curve is given by $p = f(Q, \lambda) + \gamma\varepsilon$ where $f_\lambda > 0$. Consider the effects of a change in λ. Since our equilibrium conditions are drawn in the (q, μ) plane, any parameter which only affects μ will not affect either the $H = b$ or the $G = 0$ curves. Hence, it follows immediately that an increase in λ will leave q and μ unchanged.

This neutralization of the effect of λ is accommodated by the entry of new firms into the market. Clearly, for μ to remain constant, we have

$$(17) \quad \partial f/\partial Q \cdot \partial Q/\partial\lambda + \partial f/\partial\lambda = 0,$$

so that

$$(18) \quad \partial Q/\partial\lambda = -(\partial f/\partial\lambda)/\partial f/\partial Q > 0.$$

Once again the result concerning the individual firm's output should be contrasted with Sandmo's result that (in the single-firm context) such a shift in demand will raise the individual firm's output if absolute risk aversion is constant or decreasing, but may reduce it if absolute risk aversion is increasing.

PROPOSITION 3: (a) *A mean-preserving increase in price uncertainty reduces industry output. This result is independent of absolute and/or relative risk aversion.* (b) *The effect of a mean-preserving increase in price uncertainty on the output of individual firms and on the number of firms in the industry appears to be ambiguous. This is so even if decreasing absolute and/or increasing relative risk aversion are assumed.*

PROOF:

We have from (8)

$$(19) \quad \left.\frac{d\mu}{d\gamma}\right|_{\substack{H=b \\ q=q^*}} = -\frac{H_\gamma}{H_\mu} = -\frac{E[U'(\cdot)\varepsilon]}{E[U'(\ )]} > 0$$

so that, since the $H = b$ curve shifts upwards, then regardless of how the $G = 0$ curve shifts, μ must rise and hence industry output must decline. However, from (7) we have

$$(20) \quad \left.\frac{d\mu}{d\gamma}\right|_{\substack{G=0 \\ q=q^*}} = -\frac{G_\gamma}{G_\mu}$$

$$= -\frac{qE[U''(\cdot)(p-C')\varepsilon] + E[U'(\cdot)\varepsilon]}{qE[U''(\cdot)(p-C')] + E[U'(\cdot)]}$$

which cannot be signed, unless decreasing absolute risk aversion is assumed in which case $-G_\gamma/G_\mu$ can be shown to be positive. Nonetheless, even having made this assumption it appears to be impossible to sign the difference between the shifts in the $H = b$ and $G = 0$ curves.

PROPOSITION 4: *An increase in the reservation utility will reduce industry output and raise mean price. If absolute risk aversion is nonincreasing, the individual firm's output will rise and the number of firms in the industry will fall. If absolute risk aversion is increasing, it is possible that the individual firm's output will fall and the effect on the number of firms will be ambiguous.*

PROOF:

Clearly, a change in b does not alter the $G = 0$ curve. The $H = b$ curve, however, shifts upwards.

$$(21) \quad \left.\frac{d\mu}{db}\right|_{\substack{H=b \\ q=q^*}} = -(1/H_\mu) > 0.$$

Thus, if the situation is as described in Figure 1a, it is clear that both μ and q increase. If, however, the situation corresponds to Figure 1b, then μ increases but q falls.

Finally, it may be of interest to note the effects of changes in a proportional profits tax (with full loss offset). The analysis of the effects of such a change on the single firm is given in Sandmo and extended in Katz (1983). While space considerations preclude the presentation of the long-run analysis, it appears that it is not possible to sign the long-run effect of a proportional profits tax

on any of the variables under consideration. This is the case even if we are prepared to make the assumptions of decreasing absolute and/or increasing relative risk aversion, unless an approximation of the utility function is made (see Appelbaum and Chin Lim, 1982) in which case (with these assumptions) industry output will rise with the tax rate, but the effect of the tax on q and n will remain indeterminate.

REFERENCES

Appelbaum, Elie and Lim, Chin, "Long Run Industry Equilibrium with Uncertainty," *Economic Letters*, November 1982, 139–45.

Batra, Raveendra and Ullah, Aman, "Competitive Firm and the Theory of Input Demand Under Price Uncertainty," *Journal of Political Economy*, May/June 1974, *82*, 537–48.

Katz, Eliakim, "Relative Risk Aversion in Comparative Statics," *American Economic Review*, March 1983, *73*, 452–53.

______, "The Optimal Location of the Competitive Firm Under Price Uncertainty," *Journal of Urban Economics*, July 1984, *16*, 65–75.

______, **Paroush, Jacob and Kahana, Nava,** "Price Uncertainty and the Price Discriminating Firm in International Trade," *International Economic Review*, June 1983, *23*, 389–400.

Leland, Hayne, "The Theory of the Firm Facing Uncertain Demand," *American Economic Review*, June 1972, *62*, 278–91.

Sandmo, Agnar, "On the Theory of the Competitive Firm under Price Uncertainty," *American Economic Review*, March 1971, *61*, 65–73.

Organizational Structure and Productivity

By RONALD WINTROBE AND ALBERT BRETON*

The analysis of economic growth and that of differences in productivity performance between firms and countries has mainly been conducted within a "growth accounting" approach (for example, Edward Denison, 1962). In that approach, productivity depends on technology, on the quantity and quality of labor and capital, and on other inputs; the growth rate of productivity, in turn, depends on the growth rates of these variables. This "explains" growth in the sense of breaking it up into its constituent parts. However, it does not explain why some firms (or countries) are able to achieve those changes in capital and knowledge, while others are not. The contribution of "organization" to productivity and to its growth is sometimes mentioned: in addition to its direct role, organization is held to be important in explaining the contribution of technological change and the accumulation of conventional inputs. It is presumably effective organization which makes possible, and ineffective organization which impedes, the rapid accumulation of capital inputs, the successful adaptation to newer technology, and the efficient use of labor.

However, there is at present no satisfactory theory which explains why some firms are "better organized" than others, or even what this means. In this paper, we approach this problem by suggesting that employers and employees invest in a particular form of capital: networks or lines of *trust* in one another. The distribution of networks within a firm is related to a number of aspects of organizational structure—for example, rate of turnover, promotion possibilities, and the availability and flexibility of noncontractual payments and other emoluments. The distribution of trust, in turn, is related to firm productivity. So the paper forges a link between measurable aspects of organization and productivity. We also suggest a new and consistent explanation for some other phenomena—for example, the well-known puzzle that productivity is procyclical and Marshall's hypothesis that firms inevitably tend to decline.

The structure of the paper is as follows. In Section I, we define trust capital, discuss its relationship to the concept of "reputation" developed by Benjamin Klein and Keith Leffler (1981) and Carl Shapiro (1983), and analyze the accumulation of trust by rational individuals. We then discuss the impact of the distribution of trust within a firm on that firm's productivity. Section II looks at some implications: we show the impact of changes in promotions, turnover, and noncontractual perquisites on productivity, and develop the other implications mentioned above.

I. The Distribution of Trust and Organizational Productivity

There are a number of circumstances within a firm which offer potential gains from trade between employer and employees and among employees. One of these arises from the limitations of formal contracts. A typical employee (A) is paid a contractual salary in exchange for supplying a loosely defined set of contractual services. His immediate superior, whom we will call V (denoting a vertical relationship to A), has contractual authority over A, but because of the incompleteness or rigidity of A's contract, and therefore of V's authority, there are numerous ways in which A could be induced to perform his tasks with greater efficiency if only A could trust V. Employee A could offer such things as extra effort, "consummate" rather than perfunctory cooperation (Oliver Williamson, 1975), more accurate in-

*Departments of Economics, University of Western Ontario, London, Ontario N6A 5C2, and University of Toronto, 150 St. George Street, Toronto, Ontario M5S 1A1, respectively. The ordering of names is random; we each contributed equally to this paper, as is true of all our jointly produced work. We thank Hal Hochman, Peter Kuhn, Chin Lim, Line Robillard, and Michel Rochon for help, comments, and suggestions.

formation, minimum red tape, circumvention of formal procedures, and so on. We shall call these services *informal* or *noncontractual* services. In exchange, V could offer A the prospect of more rapid promotions, bonuses, expense accounts, a larger budget, free travel, a nicer office, better working conditions, and so on. In this kind of exchange, subordinates attempt to achieve the objectives of their superior more effectively in exchange for promises from the superior of greater rewards from the resources at his disposal. We call these kinds of trades *efficient* trades, since, as discussed later, they tend to increase the efficiency of production.

An alternative strategy for subordinates is to attempt to obtain these resources through deception, trickery, or sabotage, rather than by cooperation, a class of behavior studied in William Niskanen (1971) and in other models which emphasize the pathology or the inefficiency of organizational behavior. Thus, employees can distort information about their performance to make it look better, rather than actually improving it (Gordon Tullock, 1965), distort information on the costs of their output in order to obtain a bigger budget, manipulate agendas and otherwise try to entrap their superiors. Strikes, work to rule, and slowdowns are other examples of this kind of behavior.

An important aspect of these kinds of behavior, which heretofore has received little attention, is that they often cannot be implemented by a single employee, but may require cooperation or agreement among employees. That is, they are the outcome of trades. This is obvious in the case of slowdowns or strikes, but it is often true of information distortion or agenda manipulation, and even of shirking. The reason is that the success of any of these strategies requires that they be concealed from superiors, or at least that (in the case of strategies such as work to rule or slowdowns) there is no single individual whom superiors can isolate as responsible. We shall call these strategies—in which subordinates obtain resources by tricking or sabotaging their superiors—*inefficient* trades.

How can either of these kinds of trades take place? There is no third-party enforcement agency within the organization which would prevent employees from cheating or reneging on their commitments to one another or to their employer. The prospect of future exchanges may, however, induce agents to forego the short-run gains from cheating, and agreements may, in Lester Telser's (1980) phrase, be "self-enforcing." Klein and Leffler discuss the conditions for this to happen in the context of markets for high-quality goods.

Shapiro, on the other hand, introduces the concept of "reputation" as an asset that must initially be built up before consumers are willing to purchase high-quality goods from a firm. The firm signals high quality by initially selling high-quality goods at the price of low-quality goods. In subsequent periods, the firm recovers this investment by charging a price premium. Similarly, our earlier study (1982), suggests the concept of *trust* as an asset which must be built up before informal services or agreements will be exchanged within an organization. There are a number of processes by which this asset may be created. For the purposes of this paper, we simply follow Shapiro and assume that employees will initially provide informal services without demanding full or any compensation. In subsequent periods, however, the price received for informal services or agreements must be large enough to yield a return on this investment in addition to covering their costs. As in Klein-Leffler or Shapiro, this premium acts as a deterrent against cheating if the present value of the returns from honoring commitments exceeds the one-shot quasi rent from cheating.

The concept of trust[1] as used here differs from Shapiro's concept of reputation in one respect: it refers to the confidence held by a particular buyer in a seller. This confidence may be different from that held by other buyers, and need not spread to them. Shapiro assumes (as do Klein-Leffler) that a firm's reputation is common knowledge among consumers. We shall assume that if an agent

[1] The word "trust" has the disadvantage that it sometimes has ethical connotations, being often confused with altruism (on this problem, see Wintrobe, 1981; 1983).

(B) invests in A's trust (i.e., accumulates the asset ${}_aT_b$), that this provides no grounds for another agent C to trust B. If B wishes C to trust him, it will be necessary to build trust separately with C.

To put it differently, there are two margins along which trust can be accumulated—the number of relationships and the amount invested in each one. Since trust is a capital good, B's desired investment in any single relationship (${}_aT_b$) will tend to be larger, the more B expects to trade with A in the future, the lower B's subjective rate of discount, the larger are B's opportunities to cheat on those trades, and the smaller the costs of signalling to A.[2] If both A and B possess opportunities to cheat each other (perhaps the typical case within an organization), A will also invest in the reciprocal asset ${}_bT_a$, and A and B may be said to form a *network*.

We now turn to the link between networks, or trust relationships within organization and productivity. Since both efficient and inefficient trades are based on trust, it follows that an increase in trust throughout the organization will not necessarily increase productivity. The crucial variable linking trust to productivity is not the absolute amount of trust, but its distribution.

To derive a simple link between the distribution of trust in a firm and productivity, we make the strong assumption that "vertical" or superior-subordinate trades, based on vertical trust or networks, are primarily efficient trades, while "horizontal" or subordinate-subordinate trades—based on horizontal trust—are primarily inefficient trades. To understand this assumption, consider first an analogy between networks within the firm and interest groups (which are also networks) in politics. In the economic theory of interest groups (for example, Mancur Olson, 1965, 1982; George Stigler, 1971; Sam Peltzman, 1976), it is assumed that producer coalitions (cartels, labor unions, farm organizations) tend to press for inefficient policies. The reason is that such groups would pay all of the costs, and realize only a small fraction of society's gain from the pursuit of efficient policies (those which raise aggregate income). But they can obtain the entire gain and pay only a small fraction of the cost of inefficient policies—those which redistribute income towards themselves at the cost of aggregate income.

Horizontal networks within a firm face a similar tradeoff between activities that benefit the firm as a whole and those that redistribute the firm's resources in ways which benefit themselves but reduce the value of the firm. So our assumption that horizontal trades are inefficient is analogous to the assumption concerning the efficiency of agreements among sellers in the economic theory of interest groups. On the other hand, it is usually assumed that a reduction in transactions costs between potential buyers and sellers increases efficiency (although this need not be true if the trade generates external diseconomies), and vertical trades, we suggest, more nearly resemble trades of this type (the superior is the buyer of the services of subordinates).

It might be objected that horizontal cooperation can *raise* productivity, as it does in Armen Alchian and Harold Demsetz' (1972) analysis of "team" production, as is claimed by those supporting the movement towards "participative" management, or as one finds in popular discussions of the efficiency of contemporary Japanese firms. Moreover, vertical trades can be inefficient: the cooperation of subordinates is useful to a shirking or embezzling boss.

However, a necessary (but not sufficient) condition for horizontal cooperation to raise productivity is that the number of subordinates be small. In an assembly line, for example, cooperation among some subordinates can only be used to slow the line down, not to speed it up. To do the latter, cooperation among everyone working on the line is required, and it is extremely unlikely that the required multilateral trades could take place. Indeed, this is the simplest rationale for a hierarchical relationship (see, for example, Williamson).

[2] Differences in education, family background, race, religion, status and so on are typically thought to imply relatively high costs of forming social or trading networks (see Janet Landa, 1981).

The importance of small numbers is often emphasized by proponents of participative management (for example, F. E. Emery and M. Emery, 1975; Louis Davis, 1982; J. F. Donnelly, 1977), who argue that a precondition for their programs to be effective is that workers must be broken up into small groups. So, at a minimum, one can rationalize our assumption concerning horizontal trades as applying to the typical large bureaucratic firm, and possibly not to firms which are organized on participative lines. The analysis of Alchian and Demsetz, on the other hand, would seem to be applicable only in these (relatively isolated) instances. And our assumption that vertical trades are predominantly productivity enhancing can be rationalized on the grounds that the number of trades required to affect vertical cooperation is always relatively small, except in the very largest firms. This is one way of explaining why classical "diseconomies of large-scale management" only seem to make their appearance in such firms.

In addition, while the requirement that numbers be small makes it possible for horizontal cooperation to enhance productivity, it by no means guarantees that result. In this connection, it is interesting that while the rhetoric of participative management promotes the benefits of cooperation, the contrast which is typically drawn is that between an authority-based and a cooperative, or trust-based, system (for example, Emery and Emery). The distinction between horizontal and vertical cooperation is not made. Most importantly, the schemes themselves appear to be as much concerned with increasing vertical trust as horizontal trust. For example, the key to the success of the Scanlon plan, according to one of its proponents, is the "linking pin person"—that the head of one team be also part of a successful working team at the next higher level (Donnelly, p. 120). We suggest that these schemes are effective to the extent that they promote vertical trust rather than horizontal trust, and that it is the formation of unwanted horizontal trust which explains why these schemes are sometimes unsuccessful, and why they have not been more widely adopted. Similarly, in the large Japanese firm, it is interesting to note that the much-heralded work group always meets with their supervisor, and that worker competition there, according to at least one Western observer, "often is decided by a worker's vertical relationship to his superiors. Great emphasis is placed on impressing superiors, on flattery, and behind-the-door deals. It increases friction among workers and makes it difficult to develop deep horizontal friendships and worker solidarity" (R. E. Cole, 1971, pp. 165–66).

The rest of our assumptions are straightforward. We assume that we can define the variables T_V and T_H, the aggregate quantities of vertical and horizontal trust capital, respectively, and that the firm's production function may be written in the usual way as

$$Q = f(K, L, T_V, T_H), \tag{1}$$

with T_V and T_H as well as the conventional K and L as arguments.

Productivity per man is[3]

$$q = \frac{Q}{L} = \frac{f(K, L, T_V, T_H)}{L}. \tag{1'}$$

The assumption that vertical trust enhances productivity, while horizontal trust reduces it, implies that $q_{TV} > 0$, $q_{TH} < 0$, where $q_{TV} \equiv \partial q / \partial T_V$, $q_{TH} \equiv \partial q / \partial_{TH}$. Finally, we assume that $\partial^2 q / \partial T_V^2 < 0$, and $\partial^2 q / \partial T_H^2 > 0$, that is, there are diminishing marginal productivity gains to vertical trust and increasing marginal productivity damages to horizontal trust. The first assumption is standard; the second may be rationalized as follows. Consider a network of employees that rests on a certain quantity of horizontal trust. That quantity leads to a given flow of inefficient informal services and therefore to a given reduction in output. To prevent such damage, superiors will use resources to monitor the inefficient or uncooperative behavior of subordinates. Now suppose T_H increases, and suppose also that the amount of

[3] The production function is not homogeneous of degree one in all its arguments including T_V and T_H. See the second assumption below.

resources used in monitoring and sanctioning are increased by the same proportion. But, for the usual reason, we must assume that the marginal productivity of resources used in monitoring diminishes as more and more of them are allocated to that task. The proposition that it gets *increasingly* costly to monitor uncooperative or disruptive behavior is a sufficient condition for increasing marginal damages from T_H.

We formulate our analysis in terms of T_H and not in terms of monitoring, because there are other sources of damages from an increase in T_H apart from its effect on monitoring costs. Many inefficient agreements, such as those resulting in slowdowns or strikes, cause direct losses in output. $\partial^2 q/\partial T_H^2$ would be positive so long as these marginal "direct" damages are nondiminishing. This assumption seems reasonable. Indeed, it is commonly assumed in the analysis of activities which generate external diseconomies such as crime (Gary Becker, 1968), or pollution (Richard Musgrave and Peggy Musgrave, 1976) that marginal direct damages are increasing.

II. Implications: Organizational Structure and Productivity

This section develops a set of testable hypotheses relating productivity and its growth rate to various aspects of organizational structure. These implications basically derive from two propositions. First, organizations differ in efficiency primarily because of differences among them in the amounts of vertical and of horizontal trust. The growth rate of productivity, as is the case with conventional factors, tends to be related to the growth rates of these variables: positively related to the rate of growth of T_V and negatively related to the rate of growth of T_H. The growth rate of productivity will, however, also be related to their *levels*, because, we suggest, a high level of T_V and a low level of T_H tend to permit a relatively rapid accumulation of conventional inputs —physical and human capital. For example, a major obstacle to employee investments in specific capital is the employee's mistrust of his employer (low T_V).[4] Similarly, the growth rate of physical capital and the rate of adaptation to changes in technology depend on the capacity of the organization to absorb changes in machinery and methods for using it. The higher is T_V, the greater the willingness of employees to absorb these changes. The smaller the amount of T_H, the smaller their capacity to disrupt or block them if they choose to do so.

Our second proposition is that differences in the levels of T_V and T_H can be traced to various characteristics of organizational structure. These include the amount of turnover (X), the probability of promotion (P), and the amount of noncontractual perquisites (PE). These variables are readily measurable. So, one way to test the theory is to correlate productivity—the basic dependent variable, measured by average productivity per man (q), against those variables that affect the amount and distribution of trust.

Of course, if the theory developed here is correct, all firms will want to choose an organizational structure which attempts to maximize T_V and minimize T_H. However, many variations in the structural variables considered here are outside the firm's control. Turnover rates are partly related to characteristics of the labor force and to policies of competing firms; promotions depend on the growth rate of demand. Consequently, although the best organizational structure is the one that simply maximizes T_V and minimizes T_H, the efficient organizational structure takes into account the costs involved in using different instruments for this purpose.

We shall not attempt to develop a complete model of the firm here. A complete model would describe the equilibrium organizational structure (levels of P, PE, and X) as a function of the manager's objective function and entrepreneurial capacity, characteristics of the labor force, the state of demand, and the behavior of competing

[4] This point is also made by Masanori Hashimoto (1979).

firms. For the most part we simply consider a "representative" employee and show the effects of changes in each of these variables on his optimal accumulation of vertical and horizontal trust. The aggregate variables T_V and T_H are assumed to move in the same direction as their optimal values for this employee and this, via equation (1′), gives rise to the predicted relationships between structure and productivity.

The first variable to be considered is employee turnover (X). The relevant theoretical variable is a subordinate's subjective probability that he will leave the organization. So both quits and firings are included. An increase in X reduces the anticipated number of future trades between a subordinate and his superior and between him and other subordinates as well. Thus the incentive to any subordinate to accumulate either T_V or T_H is diminished, that is, $\partial T_H/\partial X$ and $\partial T_V/\partial X$ are both negative. The fall in T_V tends to reduce productivity, while the fall in T_H tends to increase it. At this level of generality, therefore, productivity may rise or fall. Before pushing the analysis further, note that this conclusion shows a flaw in the conventional treatment of turnover (for example, Boyan Jovanovic, 1979).

In the conventional literature, turnover always reduces productivity. The costs of turnover include an increase in administration costs and in expenditures on training as well as a reduction in employee incentives to accumulate firm-specific capital. Since human capital is always productivity enhancing, an increase in turnover is always detrimental to productivity. Our theory of trust can be fitted into this framework, because trust is also a form of human capital. Where our theory breaks with this literature is in emphasizing that not all of the capital accumulated by subordinates is productivity enhancing. The standard literature assumes that it is: consequently, it has no place for cliques (networks that reduce the efficiency of the firm); for informal practices such as work-to-rule, deliberate slowing down of an assembly line, or reductions in productivity in order to raise piece rates; or for top management groups who sometimes see in fast-rising subordinates a potential threat to their position and collude to remove or reduce that threat. All of these examples of inefficiency are made possible only by investments in specific (trust) capital. And all reduce productivity.

Under what kinds of circumstances would we predict that turnover reduces productivity and so justify the assumption of the labor economics literature? In our framework, whether an increase in turnover increases or decreases productivity depends on the relative sizes of $\partial q/\partial T_V$ and $\partial q/\partial T_H$. These are determined in part by the *levels* of T_V and T_H since $\partial^2 q/\partial T_V^2 < 0$, and $\partial^2 q/\partial T_H^2 > 0$, as argued in Section I.

Consider a firm in which, initially, turnover is very high, and the absolute levels of T_V and T_H, therefore, close to zero. Now, suppose a comparative static fall in the turnover rate. At the lower rate of turnover, the yield to investments in both T_V and T_H increases. If turnover is still fairly high, the net effect may be an increase in productivity. As turnover continues to decrease, however, and T_V and T_H continue to rise, diminishing returns to T_V eventually set in, as do increasing damages from T_H; eventually, the loss in output from the change in T_H outweighs the gain from T_V.

This analysis, while simple, does provide some foundation for the proposition that firms and other organizations have an inherent tendency to decline, as suggested by writers as diverse as Marshall with respect to firms, Robert Michels (1962) with respect to political parties (the so-called Iron Law of Oligarchy) and, more recently, by Olson (1982) with respect to nation states. All of these claims, have, however, been based primarily on observation, either anecdotal (Marshall), historical (Michels), or, in the case of Olson, on impressive empirical evidence.

In our analysis, organizations tend to decline because, eventually, their members come to know each other "too well" from the point of view of the organization as a whole. Put simply, the harm caused by the increase in their capacity to make self-serv-

ing agreements with each other eventually outstrips the benefits to the organization from their capacity to cooperate with their superiors. The organization can check this development by horizontal transfer of employees, by monitoring employees more closely, by reorganizing the firm, by forcing resignations, or by outright firings, but each of these instruments ultimately has its limitations in preventing the accumulation of horizontal trust.

Note that the analysis does not imply that decline need be especially rapid. In particular, although decline ultimately takes place because of excessive horizontal trust, this need not take place within a single generation if the increase in horizontal trust leads to changes in the organization's formal structure which make it easier for subsequent generations to accumulate T_H. The analysis does imply that the highest productivity among firms in an industry should be found among neither its youngest nor its oldest firms.

The second variable to be discussed is perquisites (PE) or perks—payments to employees that are not part of the employees' contractual wage: bonuses, merit pay, or expense accounts, the fabled executive washroom privileges, use of a company car or airplane, and so on. Perks also include employee capturable gains from shirking (long breaks, on-the-job leisure, etc.). There are perks to which all employees are entitled (by custom or contract), whereas others are given on a discretionary basis. The important theoretical variable for our purpose is discretionary perks.

Perks are a means by which superiors may pay subordinates for informal services. Their chief virtue, from the point of view of superiors, is flexibility: their use need not be limited by criteria set up in formal compensation schemes and they can be given or taken away on relatively short notice. There are few enough, if any, perks from subordinate to subordinate that it is legitimate to assume that, in effect, there are none. Consequently, in the analysis of the economic function of perks one can assume that the yield to T_H remains unchanged whenever the amount of discretionary perks at the disposal of superiors changes. If the perks that superiors control are used to purchase informal services spot, very little, if any T_V need exist. Under such circumstances, perks are a substitute for T_V. On the other hand, if the perks superiors control are used for intertemporal (trust-based) trade, the net yield to T_V may increase as the volume of perks increases.

In both cases the productivity of the organization will *increase* as perks increase ($\partial q/\partial PE > 0$): in the first case, because more efficient informal services are purchased spot and in the second, because T_V makes more intertemporal trades possible. In both cases, T_H and, therefore, horizontal trades are unchanged; it follows that an increase in perks will induce subordinates to substitute efficient for inefficient trades, thus raising productivity.

One implication of this is that proposals, such as the current one in the United States, to tax perks would have an adverse effect on the productivity of business, to the extent that promotions and other noncontractual payments not covered by such legislation would be insufficient to serve as substitutes for taxable perks. Moreover, there will be inequities, since the perks taxed will only be measurable perks that will be weighted towards the discretionary rewards category.

Consider promotions (P) next. If the supply of promotions increases, the net returns to any subordinate of forming trust with his superiors—who either decide on promotions or whose reports will certainly influence that decision—obviously increases as well. This implies $\partial T_V/\partial P > 0$. On the other hand, the more a subordinate expects to be promoted, the less he anticipates interacting in the future with those at his current level, ($\partial T_H/\partial P > 0$) and the more he anticipates trading with his current superiors, ($\partial T_V/\partial P > 0$ again). These three effects of a larger supply of promotions reinforce each other with respect to firm productivity. Therefore, the firm's productivity is positively related to the frequency of promotions within it.

Increasing the supply of promotions is also costly. Administration and on-the-job training costs will tend to be higher. The

optimal level of promotions—which takes these costs into account—will certainly expand when demand and therefore the size of the organization is growing. It follows that there is also a positive relationship, *ceteris paribus*, between the growth rate of demand and productivity.

One interesting implication of the positive association between growth, promotions, and productivity is that average productivity will be procyclical, as suggested by the evidence—evidence that has puzzled economists since the 1930's. In addition, our explanation of the procyclical behavior of productivity is both simple and intuitively appealing. When demand is growing, prospects for promotion are relatively good, and employees will want to maintain or deepen their relationship with their superiors, who will be deciding whom to promote.[5] On the other hand, when demand is falling, prospects for promotion are slim, and the yield to these investments is relatively small. Network investments and trading therefore tend (other things equal) to be relatively efficient in periods of growth and relatively inefficient in periods of decline. This change in trading, which will take place throughout the organization, is reflected in the data on productivity.

III. Conclusion

In this paper, we have presented an analysis of the economists's "black box"—the relationships within a firm. The essence of the analysis is that these relationships are governed by noncontractual exchanges, and that these exchanges are made possible by the accumulation of trust between the employer and his employees (vertical trust) and among the employees (horizontal trust).

Using the strong assumption that the productivity of firms depends positively on the amount of vertical trust, and negatively on horizontal trust, we derived a number of implications with respect to the effect of organizational structure on productivity. In addition, we suggested simple and intuitively appealing explanations for a number of otherwise puzzling phenomena: for example, that productivity tends to be procyclical (since growing demand tends to raise the attractiveness of vertical relative to horizontal exchange), Marshall's hypothesis that firms and other organizations tend to decline with age (as both vertical and horizontal trust accumulate, the productivity-diminishing effects of horizontal exchanges ultimately outweigh the productivity-augmenting effects of vertical exchanges), and that increased turnover need not diminish productivity.

[5]It has been suggested that when demand is falling, employees may also want to deepen their relationship with their supervisors to avoid being dismissed. However, the yield to T_V *falls* as the probability of dismissal rises (see the discussion of turnover in the text). In any case, the strategy may not be credible, and superiors can be expected to favour those who are already members of their networks, and who will at the same time be "calling in" their loans to avoid dismissal. See our earlier study for further discussion.

REFERENCES

Alchian, Armen A. and Demsetz, Harold, "Production, Information Costs, and Economic Organization," *American Economic Review*, December 1972, *62*, 777–95.

Becker, Gary S., "Crime and Punishment: An Economic Approach," *Journal of Political Economy*, March/April 1968, *76*, 169–217.

Breton, Albert and Wintrobe, Ronald, *The Logic of Bureaucratic Conduct*, New York: Cambridge University Press, 1982.

Cole, R. E., *Japanese Blue Collar: The Changing Tradition*, Berkeley: University of California Press, 1971.

Davis, Louis E., "Organization Design," in G. Sadvendy, ed., *Handbook of Industrial Engineering*, New York: Wiley & Sons, 1982.

Denison, Edward F., *The Sources of Economic Growth in the United States and the Alternatives before Us*, New York: Committee for Economic Development, 1962.

Donnelly, J. F., "Participative Management at Work: An Interview with John F. Donnelly," *Harvard Business Review*, January/February 1977.

Emery, F. E. and Emery, M., "Participative Design: Work and Community Life," Occasional Papers in Continuing Education, No. 4, Centre for Continuing Education, Australian National University, 1975.

Hashimoto, Masanori, "Bonus Payments, On-The-Job Training, and Lifetime Employment in Japan," *Journal of Political Economy*, October 1979, *87*, 1086–104.

Jovanovic, Boyan, "Job Matching and the Theory of Turnover," *Journal of Political Economy*, October 1979, *87*, 972–90.

Klein, Benjamin, and Leffler, Keith B., "The Role of Market Forces in Assuring Contractural Performance," *Journal of Political Economy*, August 1981, *89*, 615–41.

Landa, Janet T., "A Theory of the Ethnically Homogeneous Middleman Group: An Institutional Alternative to Contract Law," *Journal of Legal Studies*, June 1981, *10*, 349–62.

Michels, Robert, *Political Parties*, New York: Free Press, 1962.

Musgrave, Richard A. and Musgrave, Peggy B., *Public Finance in Theory and Practice*, 2nd ed., New York: McGraw-Hill, 1976.

Niskanen, William A., Jr., *Bureaucracy and Representative Government*, Chicago: Aldine-Atherton, 1971.

Olson, Mancur, *The Logic of Collective Action*, Cambridge: Harvard University Press, 1965.

______, *The Rise and Decline of Nations*, New Haven: Yale University Press, 1982.

Peltzman, Sam, "Towards a More General Theory of Regulation," *Journal of Law and Economics*, August 1976, *19*, 211–40.

Shapiro, Carl, "Premiums for High Quality Products as Returns to Reputations," *Quarterly Journal of Economics*, November 1983, *98*, 659–79.

Stigler, George J., "The Theory of Economic Regulation" *Bell Journal of Economics*, Spring 1971, *2*, 3–21.

Telser, Lester, "A Theory of Self-Enforcing Agreements", *Journal of Business*, January 1980, *53*, 27–40.

Tullock, Gordon, *The Politics of Bureaucracy*, Washington: Public Affairs Press, 1965.

Williamson, Oliver E., *Markets and Hierarchies*, Glencoe: Free Press, 1975.

Wintrobe, Ronald, "It Pays To Do Good, But Not To Do More Good Than It Pays," *Journal of Economic Behaviour and Organization*, September 1981, *2*, 201–13.

______, "Taxing Altruism," *Economic Inquiry*, April 1983, *21*, 255–69.

Robert Giffen and the Irish Potato: Note

By ULRICH KOHLI*

In their paper in this *Review* (1984), Gerald Dwyer and Cotton Lindsay express serious doubts about the existence of a Giffen-good paradox during the 1845–49 Irish potato famine. While many of their arguments and their main conclusion are convincing, one important step in their demonstration needs to be qualified.

Dwyer and Lindsay argue that nineteenth-century Ireland was essentially a closed economy, and hence that the supply of potatoes was exogenous. The sharp drop in the supply of potatoes in 1845–49 must therefore be viewed as a market experiment, rather than an individual experiment, with the price of potatoes an endogenous variable. Dwyer and Lindsay consider the following as a major problem with the common view that potatoes are a Giffen good: "...With an upward-sloping demand curve, a decrease in supply results in a lower, not a higher, price....This possibility is ludicrous: a large part of a country on the brink of starvation, and the market price of potatoes falls?" (p. 189). Dwyer and Lindsay then present some indirect evidence that the price of potatoes, far from falling, actually increased. This outcome seems inconsistent with the Giffen-good hypothesis.

The purpose of this note is to show with the help of a simple general equilibrium model that evidence of a price increase is not necessarily incompatible with the Giffen-good hypothesis, although the hypothesis must nevertheless be rejected in the Irish case on the basis of additional evidence presented (but not fully exploited) by Dwyer and Lindsay regarding the movement of prices of potato substitutes. This exercise proves instructive, for it reveals an interesting property of the Giffen-good model that seems to have gone unnoticed in the past.

The analysis presented by Dwyer and Lindsay is in terms of the inverse of the Marshallian demand curve. That is, they implicitly treat nominal income and the prices of all goods other than potatoes as given. For simplicity, assume that there are only two goods, and let the Marshallian demand for potatoes be given by

$$q_P = d(p_P, p_M, Y), \tag{1}$$

where q_P is the quantity of potatoes, p_P is the price of potatoes, p_M is the price of the other good (say meat), and Y is nominal income. The inverse of the Marshallian demand function expresses the price of potatoes as a function of the quantity available, and is defined as

$$p_P = g(q_P, p_M, Y). \tag{2}$$

One can easily see that $\partial g(.)/\partial q_P = [\partial d(.)/\partial p_P]^{-1}$. It is well known that $\partial d(.)/\partial p_P > 0$ if potatoes are a Giffen good, hence $\partial g(.)/\partial q_P > 0$ as well. That is, a decrease in the available supply of potatoes must lead to a reduction in the price of potatoes. As noted by Dwyer-Lindsay, any evidence to the contrary should lead to the rejection of the Giffen-good hypothesis.

Crucial in the above argument is the assumption that p_M is given. While Ireland was a closed economy in potatoes, it was not isolated economically, and it may well be that other goods were available at given international prices. This is possibly the view implicitly held by Dwyer and Lindsay. However, there is a certain ambiguity in this regard, for they explicitly argue that Ireland was a peasant economy that was closed in food. Furthermore, they present evidence that the prices of potato substitutes increased substantially during the famine. This last

*Professor of Economics, University of Geneva, 2 Rue de Candolle, CH–1211 Geneva 4, Switzerland. I am grateful to W. Erwin Diewert, Philip Graves, Edward Morey, and an anonymous referee for their comments on an earlier version of this note.

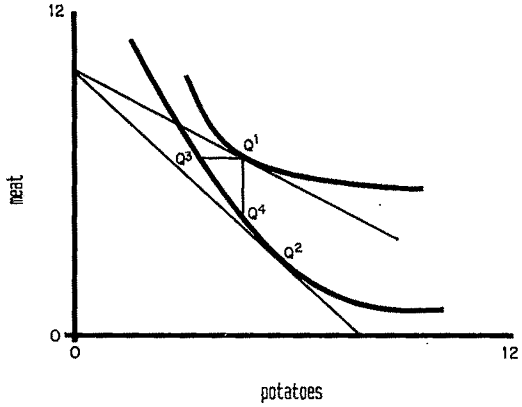


FIGURE 1

observation is inconsistent with the assumption of a perfectly elastic supply.

Let us assume therefore that rural Ireland was indeed a closed economy, that is, that the supplies of potatoes as well as of meat were given. The price effect of an exogenous reduction in the supply of potatoes can then be analyzed with the help of the inverse demand for potatoes. Inverse demand schedules express prices as functions of quantities.[1] In this example, the inverse demand for potatoes is defined as

$$p_P = h(q_P, q_M, Y), \tag{3}$$

where q_M is the available quantity of meat. Note the difference between (2) and (3).[2] In the former, the price of meat is considered exogenous and the quantity demanded is implicitly treated as endogenous, while the reverse is true in the latter.

Inverse demand functions are generally downward sloping (for example, see Ronald Anderson), that is, $\partial h(.)/\partial q_P < 0$, and it turns out that this is also true for Giffen goods. This can be shown graphically with Figure 1 that depicts a pair of indifference curves in the quantity (potatoes, meat) space. The graph is drawn under the assumption that potatoes are a Giffen good at point Q^1. (An increase in the price of potatoes, as depicted by the steeper budget line, would indeed raise the demand for potatoes, with equilibrium moving from Q^1 to Q^2.)

Figure 1 can easily be used for market experiments. Once the supply of goods is given, the marginal rate of substitution at that point determines the market-clearing price ratio. Knowing nominal income, absolute prices can be determined as well. Consider now the effect of a drop in the supply of potatoes, for a given quantity of meat. This change, described in Figure 1 by a movement from Q^1 to Q^3, is accompanied by an increase in the marginal rate of substitution between meat and potatoes; that is, the relative price of potatoes increases. This result is hardly surprising: a drop in the supply of potatoes, by reducing real income, raises the demand for potatoes and lowers the demand for meat. Moreover, it is visible from Figure 1 that, compared to the original situation, the vertical intercept of the price line (new position not drawn) moves upwards, while the horizontal intercept shifts to the left. Hence, for given money income, the price of potatoes increases and the price of meat falls in absolute terms.[3]

The analysis can also be conducted in terms of the Marshallian demand curve. Figure 2 is similar to Dwyer and Lindsay's graph (lower panel). Dwyer and Lindsay argue that a decrease in the supply of potatoes, from q_P^1 to q_P^2, leads to a price fall, from p_P^1 to p_P^2, under the Giffen-good hypothesis. Implicit in their analysis is the assumption that the price of meat remains unchanged. Yet if the supply of meat is given, its price must fall as the result of the reduction in the supply of potatoes. Hence the Marshallian demand curve must shift

[1] I use the terminology of Ronald Anderson (1980), W. Erwin Diewert (1980), and John Weymark (1980).

[2] Many authors, when discussing inverse demand, do so in a one-good context, in which case the distinction between (2) and (3) vanishes. See Hal Varian (1978), and James Henderson and Richard Quandt (1971, p. 134), for instance.

[3] See my 1985 paper for a mathematical proof of the same proposition.

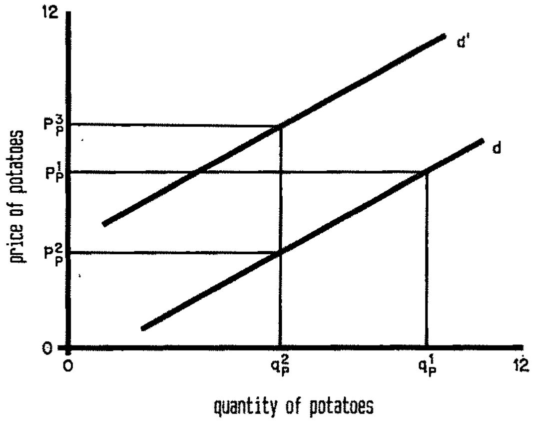


FIGURE 2

leftwards, from d to d'. The earlier discussion points out that the shift must be large enough to result in an increase in the price of potatoes, from p_P^1 to, say, p_P^3.

In these circumstances, an increase in the price of potatoes is not inconsistent with the Giffen-good hypothesis. What contradicts the hypothesis, however, is the evidence of an increase in the price of potato substitutes. Indeed, the model predicts that, for given q_M and Y, a reduction in q_P results in a fall in p_M if potatoes are a Giffen good. Although nominal income almost certainly did not remain constant during the 1845–49 events, it is hard to imagine that it actually increased. Hence a rise in p_M is incompatible with the Giffen-good hypothesis. Thus, Dwyer and Lindsay's conclusion remains intact, but the relevant piece of evidence is the increase in the price of potato substitutes, not the increase in the price of potatoes.

Dwyer and Lindsay conclude their discussion by arguing that the Giffen-good concept has no application in closed economies where the supplies of goods are given. This is not necessarily true. Giffen goods may exist in open as well as in closed economies, but it simply turns out that Giffen goods do not display any spectacular property with respect to inverse demands. This is not so for meat, however. It is worthwhile noting that a decrease in the supply of meat (for example, resulting from an outbreak of foot-and-mouth disease, depicted by the movement from Q^1 to Q^4 in Figure 1) would actually result in a *lower* price of meat. This is visible from the upward movement of the intercept of the price line when the endowments point shifts from Q^1 to Q^4. Hence, for given nominal income, the price of meat must fall. That is, the uncompensated inverse demand for meat is positively sloped,[4] and meat could therefore be termed an *anti-Giffen* good.[5] This property of the potatoes-meat model seems to have gone unnoticed in the past, for attention is generally focused on potatoes, yet is is meat that behaves abnormally when it comes to inverse demand.

[4] The uncompensated inverse price elasticity can be decomposed into two parts: a pure substitution term and a scale effect (Anderson). Although scale effects are normally negative, it is positive in the case of meat; moreover, it is large enough to dominate the negative substitution effect. Put another way, the income elasticity of meat is so large that a drop in its supply, by lowering real income, results *ceteris paribus* in a fall in demand that actually exceeds the reduction in supply.

[5] This name is inspired by I. F. Pearce's (1964) use of the term antidemand to designate inverse demand analysis. Jack Hirshleifer (1980) uses the term ultrasuperior to refer, in the two-dimensional case, to the partner of an inferior good. Anti-Giffen goods are ultrasuperior, but the reverse is not necessarily true.

REFERENCES

Anderson, Ronald W., "Some Theory of Inverse Demand for Applied Demand Analysis," *European Economic Review*, November 1980, *14*, 281–90.

Diewert, W. Erwin, "Duality Approaches to Microeconomic Theory," in K. F. Arrow and M. D. Intriligator, eds., *Handbook of Mathematical Economics*, Vol 2, Amsterdam: North-Holland, 1980.

Dwyer, Gerald P., Jr. and Lindsay, Cotton M., "Robert Giffen and the Irish Potato," *American Economic Review*, March 1984, *74*, 188–92.

Henderson, James M. and Quandt, Richard E., *Microeconomic Theory*, 2nd ed., New York: McGraw-Hill, 1971.

Hirshleifer, Jack, *Price Theory and Applications*, Englewood Cliffs: Prentice-Hall, 1980.

Kohli, Ulrich, "Inverse Demand and Anti-Giffen Goods," *European Economic Review*, April 1985, *27*, 397–404.

Pearce, Ivor F., *A Contribution to Demand Analysis*, Oxford: Oxford University Press, 1964.

Varian, Hal., *Microeconomic Analysis*, New York: W. W. Norton, 1978.

Weymark, John A., "Duality Results in Demand Theory," *European Economic Review*, November 1980, *14*, 377–95.

Marx and Malthusianism: Comment

By MIGUEL D. RAMIREZ*

In a recent article in this *Review* (1984), Samuel Hollander suggests that the secular path of real wages in the Marxian model tends toward a subsistence wage at which population growth ceases. He attributes this decline to a higher growth rate in population relative to a positive but decreasing rate of growth in the demand for labor power. From this he contends that, in contrast to the Malthusian prescription, Marx "...is open to the objection that, with no check at all to the population growth rate, the deterioration would have been sharper still" (p. 148). The implication of Hollander's argument is clear enough, for if accepted, it would necessarily lead one to dismiss Marx's claim that: "...every special historic mode of production has its own special laws of population, historically valid within its limits alone. An abstract law of population exists for plants and animals only, and only in so far as man has not interfered with them (*Capital*, I, 1967, p. 632).

This comment takes issue with the thesis advanced by Hollander on two major points. First, the secular decline in the value of labor power is a result of the increasing productivity of labor rather than the divergence between the respective growth rates in population and the demand for labor power. Second, this is perfectly consistent with constant or even rising absolute real wages (price of labor power) for the active part of the working class.[1]

Thus, population control is neither a necessary nor a sufficient condition in assuring against falling real wages for a changing social productivity of labor. Put differently, the supply of labor power to the advanced capitalist sector is the crucial supply variable —not the rate of population growth.

I begin by reviewing Hollander's useful distinction between Marx's value of labor power and the classical school's minimum subsistence wage. Next, it is shown that the value of labor power is determined by the social productivity of labor rather than factors exogenous to the system (population). The discussion is brought to a close by focusing on the prime mover of the path of relative real wages in the Marxian model: the endogenously determined supply of labor power (surplus population) to the advanced capital sector.

I. The Minimum Limit and the Value of Labor Power

Hollander correctly identifies the relationship between the value of labor power and the subsistence wage of the classical school. However, rather than quoting at length from Marx's speech on "Wages, Price and Profit" to the workers in 1865, he could have referred directly to *Capital* for a more complete and formal presentation of the subject. In chapter VI, Marx observes that

> The minimum limit of the value of labour-power is determined by the value of the commodities, without the daily supply of which the labourer cannot renew his vital energy, consequently by the value of those means of subsistence that are physically indis-

*Department of Economics, University of North Florida, Jacksonville, FL 32216. I acknowledge the valuable comments and criticisms of Cindy A. Jacobs.

[1] Ernest Mandel observes that "The idea that the real wages of the workers decline more and more is totally alien to Marx's writings" (1968, p. 151). He goes on to say, "What one finds in Marx is an idea of the absolute impoverishment not of the workers, the wage earners, but of that section of the proletariat which the capitalist system throws out of the production process: unemployed, old people, disabled persons, cripples, the sick, etc." (p. 151). A similar but more formal argument is advanced by Thomas Sowell (1960) in his seminal article. Lastly, William Baumol (1983) makes an eloquent case for Marx's rejection of the Malthusian model (see pp. 305–06).

> pensable. If the price of labour-power falls to this minimum, it falls below its value, since under such circumstances it can be maintained and developed only in a crippled state. But the value of every commodity is determined by the labour-time requisite to turn it out so as to be of normal quality.
> [Vol. I, p. 173]

By "normal quality," I believe Marx means that this particular commodity must be supplied to the capitalists in a state which ensures the latter's continuous conversion of money into capital. In his words, "...the sum of the means of subsistence...must include the means necessary for the labourer's substitutes, i.e., his children, in order that this race of peculiar commodity-owners may perpetuate its appearance in the market" (*Capital*, I, p. 172).

Two points must be made here. First, the value of labor power presupposes a positive rate of population growth. More precisely, if we denote the value of labor power by v, the minimum limit by m, and the fixed differential between them by $\bar{d}$, then the rate of change in population is given by

$$(1) \qquad dP/dt = a\bar{d}(t), \qquad a>0$$

$$\text{and} \qquad dP/dt>0$$

since $\bar{d}(t)=(v(t)-m(t))>0$.

The value of labor power and the subsistence wage are posited as functions of time for, as shall be seen in the next section, they vary with the social productivity of labor in the wage-goods industries. Second, the price of labor power would never be reduced to its minimum limit for any significant period of time since that would undermine the accumulation process itself. As Marx correctly observes, "Then again, the labour-power itself must be of average efficacy. In the trade in which it is being employed, it must possess the average skill, handiness and quickness prevalent in that trade, and our capitalist took good care to buy labour-power of such normal goodness" (*Capital*, I, p. 196). Hollander is clearly in error to suggest that Marx believed that the price of labor power would be reduced to its minimum limit.

II. The Secular Path of Wages and Technological Change

The one outstanding feature of the capitalist system is its immense social productivity. It is ceaselessly changing the methods of production so as to permit ever-increasing rates of accumulation. Marx, as opposed to Malthus and Ricardo, clearly understood that technological change was a necessary condition for the continued existence of the capitalist system as the following passage reveals:[2] "...with the advance of capitalist production and the attendant development of the productiveness of social labour and multiplication of the production branches, hence products, the same amount of value represents a progressively increasing mass of use-values and enjoyments" (*Capital*, III, p. 219).

This ever-increasing quantity of "use-values and enjoyments" would not only benefit the capitalist class, but also the active part of the working class since, according to Marx, "...it is possible with an increasing productivity of labor, for the price of labour-power to keep on falling, yet this fall to be accompanied by a constant growth in the mass of the labourers means of subsistence" (*Capital*, I, p. 523).

Marx goes on to make a similar, yet more direct statement between productivity and absolute real wages later: "But hand-in-hand with the increasing productivity of labour, goes, as we have seen, the cheapening of the labourer, therefore a higher rate of surplus value, even when the real wages are rising. The latter never rise proportionately to the productive power of labour" (*Capital*, I, p. 604).

Finally, Marx remarks in Volume III of *Capital* that "If wages fall in consequence of

[2] For a simple, but not simplistic, explanation of the process of accumulation see Paul Sweezy (1970, pp. 75–95). Fred Gottheil also provides a lucid account of the role of technological change in the Marxian model (1966, pp. 88–115).

a depreciation in the value of labour-power (which may even be attended by a rise in the real price of labour), a portion of the capital hitherto invested in wages is released" (p. 114). Thus, Hollander's assertion that Marx held the view that real earnings declined in the ordinary sense of that term is completely unfounded (see p. 146).[3] In fact, Marx was well aware of the relative deterioration of the worker's position since, he remarks,[4] "But even in such case, the fall in the value of labour-power would cause a corresponding rise of surplus-value, and the abyss between the labourer's position and that of the capitalist would keep widening" (*Capital*, I, p. 523).

Hollander's thesis, then, is reduced to a presumed deterioration in *relative* real wages as a result of "...a greater proportional decline in the growth rate of labor demand than supply..." (p. 148). Let us examine it.

III. Surplus Population and the Path of Relative Real Wages

The key supply variable in the Marxian model is not the rate of population growth, as Hollander presumes, but the supply of a surplus population to the capitalist sector. It is inextricably connected with the development of the productive forces[5] which ceaselessly takes place under this form of production. Marx was quite unambiguous on this point:

> But if a surplus labouring population is a necessary product of accumulation or of the development of wealth on a capitalist basis, this surplus-population becomes...a condition of existence of the capitalist mode of production. It forms a disposable industrial reserve army, that belongs to capital quite as absolutely as if the latter bred it at its own cost. Independently of the actual increase of population.... [*Capital*, I, p. 632]

The nature of this surplus population—and its secular rise—can only be understood by referring to its three major components: what Marx terms the floating, the latent, and the stagnant forms of the relative surplus population. The first of these forms is generated by "...automatic factories...where machinery enters as a factor, or only the modern division of labor is carried out..." (*Capital*, I, p. 641). That is, the accumulation of capital not only increases the demand for labor, it also displaces labor via its introduction of capital-intensive technology. In Marx's words,

> ...its accumulation, on the one hand increases the demand for labour, it increases on the other the supply of labourers by the "setting free" of them, whilst at the same time the pressure of the unemployed compels those that are employed to furnish more labour, and therefore makes the supply of labour, to a certain extent, independent of the supply of labourers. [*Capital*, I, p. 640]

The latent component is generated from the extension of capitalist production to the

[3] In a similar vein, the value of labor power falls by less than the increase in the productivity of labor in the wage-goods industries. More precisely,

$$\dot{v} = \alpha\dot{e}, \qquad 1 > |\alpha| > 0$$

where $\dot{e}$ is the rate of growth in productivity in the wage-goods industires. This arises because with the growth in the social productivity of labor, many goods which were once luxury goods are now incorporated in the subsistence minimum. For further detail, see Sowell's argument (pp. 115–16). Also, Mandel observes that "...the growth in the productivity of labour has a contradictory effect on wages. To the extent that it reduces the value of the means of subsistence it tends to cut down...the value of labour-power. To the extent that it reduces the value...of many luxury products...it tends, on the contrary to increase the value of labour-power" (pp. 147–48).

[4] A more direct statement of this view is found in Marx's address to the First International (1865) where he declares that "By virtue of the increased productivity of labour...the value of labour would have sunk, but that diminished value would command the same amount of commodities as before.... Although the labourer's absolute standard of life would have remained the same, his relative social position, as compared with that of the capitalist, would have been lowered" (*Wages, Price and Profit*, p. 65).

[5] Marx correctly assumes that technological change is capital intensive (labor saving) in nature. For further detail, see *Capital*, I, pp. 621–28.

agricultural sector. For as soon as capital-intensive methods of production are introduced in agriculture: "...the demand for an agricultural labouring population falls absolutely.... Part of the agricultural population is therefore constantly on the point of passing into an urban or manufacturing proletariat.... This source of relative surplus-population is thus constantly flowing" (*Capital*, I, p. 642).

Finally, we have the stagnant form of the relative surplus population, viz., that component which "...forms a part of the active labour army, but with extremely irregular employment" (p. 643). According to Marx, it is characterized by both relatively longer hours of work and lower real wages. Its recruits are by and large "...from those decaying branches of industry where handicraft is yielding to manufacture, manufacture to machinery" (p. 643).

There exists, therefore, a continuum with the precapitalist sector, agriculture, at one extreme, and the advanced capitalist sector, automated factories, at the other. The intermediate range is composed of family businesses, mining enterprises, small-scale manufacture and industry. At this juncture, Hollander argues that with the increasing rate of accumulation, the net demand for labor in the advanced industries of the capitalist sector increases at a decreasing rate. He goes on to say that this presumes an absolute increase in population; and thereby concludes that a reduced population growth rate can retard or prevent falling relative real wages. Meanwhile, he has overlooked the fact that the introduction of capital-intensive methods of production in agriculture and the decaying part of the capitalist sector is constantly increasing the supply of labor power to the advanced part.

From this it follows that even with a deceleration in the population growth rate, if it is not sufficiently great to offset the rising growth rate in the surplus population, it will not prevent or retard *relative* real wages from falling. That this is what Marx had in mind is evident from the following: "What were the consequences for the Irish labourers left behind and freed from the surplus-population? That the relative surplus-population is today as great as before 1846; that wages are just as low.... The revolution in agriculture has kept pace with the absolute depopulation" (*Capital*, I, p. 704).

Hollander dismisses this argument against Malthus as unconvincing since it is the result of "The intervention of 'disturbing causes,' in the present case the agricultural revolution..." (p. 149). However, what he—and Malthus before him—fail to realize is that the revolution in agriculture arises not in spite of, but because of the capitalist mode of production. It is not a fortuitous occurrence, but a necessary condition for the continued existence of capitalist production. And this brings us to the fundamental difference between Marx and the classical school, for as Paul Sweezy remarked:

> The latter is, in principle, unconcerned with changes in methods of production; economic development is viewed exclusively in terms of (gradual) quantitative changes in population, capital, wages, profit, and rent. Social relations remain unaffected; the end product is simply a state of affairs in which all these rates of change equal zero. Since the Marxian view lays primary stress on changes in methods of production, it implies qualitative change in social organization and social relations as well as quantitative change in economic variables as such. [1970, p. 94]

IV. Conclusion

Three major conclusions have emerged from this comment. First, the price of labor power would never be reduced to its minimum limit since that would threaten the very existence of the capitalist system. Second, Hollander incorrectly attributes to Marx a theory of wages which calls for a fall in absolute real wages. Finally, the crucial supply variable is the endogenously determined surplus population rather than the total population.

REFERENCES

Baumol, William J., "Marx and the Iron Law of Wages," *American Economic Review*

Proceedings, May 1983, *73*, 303–08.

Gottheil, Fred M., *Marx's Economic Predictions*, Evanston: Northwestern University Press, 1966.

Hollander, Samuel, "Marx and Malthusianism," *American Economic Review*, March 1984, *74*, 139–59.

Mandel, Ernest, *Marxist Economic Theory*, Vol. I, New York: Monthly Review Press, 1968.

Marx, Karl, *Wages, Price and Profit*, Peking: Foreign Languages Press, 1975.

______, *Capital*, New York: International Publishers, Vols. I–III, 1967.

Sowell, Thomas, "Marx's 'Increasing Misery' Doctrine," *American Economic Review*, March 1960, *50*, 111–20.

Sweezy, Paul M., *The Theory of Capitalist Development*, New York: Monthly Review Press, 1970.

Marx and Malthusianism: Reply

By SAMUEL HOLLANDER*

There is a singular unwillingness to recognize Marx's pronouncements regarding the tendency of the commodity wage to decline, the subject matter of my 1984 paper. Thus Miguel Ramirez (1986) concedes only a secular decline in the value of labor power reflecting increasing productivity. He also allows a *relative* decline in real wages compared with the return to property, which he accounts for by an inflow of labor into the advanced capitalist sector from agriculture and "the decaying part of the capitalist sector" (p. 546)—rather than by population growth. I do not see that Ramirez has demonstrated the inadequacy of the textual evidence used in support of my position that Marx intended a decline in the commodity wage. Nonetheless, he raises the valid question whether it is in fact the inflow into the advanced industrial sector rather than net population growth in the face of a deceleration in labor demand that is responsible for the secular fall in commodity wages as I maintain in my earlier paper.

One of the "laws" of capitalism is the conversion of the working class "into a class...whose ordinary wages suffice, not only for its maintenance, but for its *increase*" (*Capital*, I, p. 581; emphasis added). I underscore the argument by calling attention to the particularly high death rate among factory operatives: "The consumption of labour-power by capital is...so rapid that the labourer, half-way through his life, has already more or less completely lived himself out.... It is precisely among the work people of modern industry that we meet with the shortest duration of life" (p. 641). Marx, taking for granted net population growth as an aspect of the general process of capitalist development, points to peculiarly high marriage and birth rates to assure such growth:

> In order to conform to these circumstances [physical disability and high mortality] the absolute increase of this section of the proletariat must take place under conditions that shall swell their numbers, although the individual elements are used up rapidly. Hence, rapid renewal of the generations of labourers (this law does not hold for the other classes of the population). This social need is met by early marriages, a necessary consequence of the conditions in which the labourers of modern industry live, and by the premium that the exploitation of children sets on their production. [p. 642]

The net increase in labor supply in the modern industrial sector is thus a consequence of *internal* population growth, for which reason my original discussion was limited to that sector (see p. 140, fn. 2).

Both Ramirez and I have neglected the role of juvenile labor. A reference to the employment of child labor (made possible by "improvements in machinery") causing a "fresh disturbance to the rate of wages" implies a *supplementary* force to those I have been elaborating (*Capital*, I, p. 433; cf. pp. 394, 406–07, 472). On the other hand, a statement that the requirements of capital accumulation entail "larger numbers of youthful labourers, a smaller number of adults" (p. 641), suggests a possible *alternative* reading of Marx's declaration regarding the inconceivability that the "entire nation accomplish its total production in a shorter time span" (*Capital*, III, p. 258; my paper, p. 144). For this raises the possibility that the adult wage, in Marx's view, is reduced in consequence of an *absolute fall* in labor demand, albeit more than compensated for by an *increase* in the demand for juvenile labor.

*Department of Economics, University of Toronto, Toronto, Canada M5S 1A1.

(Women's labor and displacement by unskilled labor raise further issues.) Yet even were this the case, the countervailing potential of population control cannot be neglected; for the increased demand for child labor is met by a high birth rate, a reduction of which would surely reduce the attractiveness of this source of supply and lessen the downward pressure on the adult wage.

What now of movements into the advanced sector by displaced agricultural workers (for example, *Capital*, I, p. 708; and p. 681n regarding the inflow into mining)? While Ramirez does well to raise the question, the secular fall in the real industrial wage cannot be explained by sectoral transfer.

Consider again Marx's declaration that "a development of productive forces which would diminish the absolute number of labourers, i.e., enable the entire nation to accomplish its total production in a shorter time span, would cause a revolution..." (*Capital*, III, p. 258). This means that the net increase in labor demand in the advanced industrial sector exceeds the net decrease in agriculture and elsewhere. Under these conditions even were the entire displaced agricultural labour force to flow into the modern industrial sector, there should be no downward pressure on the real wage—*unless* there is some further source supplementing net labor supply as I have argued there is.

But there is a further matter. Marx seems to reason as if such transfers as occur are largely responses to cyclical peaks of industrial activity. A flow from agriculture then at the most puts a damper on the extent industrial wages can rise cyclically. I have already shown (p. 147) that the function of the Industrial Reserve Army is to provide "the possibility of throwing great masses of men suddenly on the decisive points without injury to the scale of production in other spheres" (*Capital*, I, p. 632). Agricultural labor would doubtless provide one source (cf. pp. 662–63). Moreover, part of the outflow from agriculture is to nonindustrial urban occupations—in which case it would have little if any effect on industrial wages. These propositions are also supported by the following:

> As soon as capitalist production takes possession of agriculture, and in proportion to the extent to which it does so, the demand for an agricultural labouring population falls absolutely, while the accumulation of capital employed in agriculture advances, without this repulsion being, as in non-agricultural industries, compensated by a greater attraction. Part of the agricultural population is therefore constantly on the point of passing into an urban or manufacturing proletariat, and *on the look-out for circumstances favourable to this transformation*. (Manufacture is used here in the sense of all non-agricultural industries.) This source of relative surplus-population is thus constantly flowing. But the constant flow towards the towns presupposes in the country itself, a constant latent surplus-population, the extent of which becomes evident only when its channels of outlet open to *exceptional width*. The agricultural labourer is therefore reduced to the minimum of wages, and always stands with one foot already in the swamp of pauperism.
>
> [*Capital*, I, p. 642; emphasis added]

Marx goes out of his way to emphasize the attraction provided by the nonindustrial urban sector, and the likely destiny of the agricultural emigrant within the town's "pauper" population.

The "domestic industry" segment of the urban population, for its part, is indeed said to provide capital with "an inexhaustible reservoir of disposable labour-power" (p. 643). Yet "its conditions of life sink below the average normal level of the working-class; this makes it at once the broad basis of special branches of capitalist exploitation. It is characterised by maximum of working-time, and minimum of wages" (p. 643). Again, Marx gives the impression that, at most, the advanced industrial sector draws upon this sector for its exceptional needs—otherwise it is incomprehensible that the wage differential should not be eradicated.

As for secular trends, it is helpful to think of the nonindustrial urban sector as im-

pinged upon by tendencies in the advanced industrial sector (and in agriculture) rather than the reverse: "It recruits itself constantly from the supernumerary forces of modern industry and agriculture, and specially from those decaying branches of industry where handicraft is yielding to manufacture, manufacture to machinery. Its extent grows, as with the extent and energy of accumulation, the creation of a surplus-population advances" (*Capital*, I, p. 643).

It remains to add that net population growth is actually a feature not only of the advanced industrial sector but of agriculture, too—over the decade 1851–61 the rural growth rate was recorded by the Census as 6.5 percent, the difference with the 17.3 percent in the towns ascribed to migration from the country (*Capital*, I, p. 642n). The "domestic industry" segment of the urban population is also said to be "a self-reproducing and self-perpetuating element of the working-class, taking a proportionally greater part in the general increase of that class than the other elements" (p. 643). It is in this context (though possibly not restricted thereto) that we encounter the famous declaration:

> [N]ot only the number of births and deaths, but the absolute size of the families stand in inverse proportion to the height of wages, and therefore to the amount of means of subsistence of which the different categories of labourers dispose. This law of capitalistic society would sound absurd to savages, or even civilised colonists. It calls to mind the boundless reproduction of animals individually weak and constantly hunted down.
>
> [*Capital*, I, p. 643]

Thus even were it the case that Marx's downward secular trend in the advance industrial sector turned on an inflow from other sectors, it would still be impossible to neglect the demographic component of the analysis.

REFERENCES

Hollander, Samuel, "Marx and Malthusianism: Marx's Secular Path of Wages," *American Economic Review*, March 1984, *74*, 139–59.

Ramirez, Miguel D., "Marx and Malthusianism: Comment," *American Economic Review*, June 1986, *76*, 543–47.

Marx, Karl, *Capital*, Moscow: Foreign Languages Publishing House, Vol. I, 1954; Vol. III, 1962.

Labor Supply and Tax Rates: Comment

By DAVID M. BETSON AND DAVID GREENBERG*

In a recent paper in this *Review* (1983) James Gwartney and Richard Stroup (G-S) take issue with what they assert is the "widespread view" (p. 447) that because changes in tax rates have a theoretically indeterminate effect on individual labor supply, the impact of changing tax rates is also indeterminate at the aggregate level. Gwartney and Stroup suggest, for example, that if the initial output of public goods is optimal and the government cuts taxes and expenditures by equal amounts, the forgone public goods would be valued as highly as the private goods that individuals purchase with their increase in disposable income. Hence, aggregate real income would not rise, but remain at its pre-tax-cut level. Gwartney and Stroup state that under these circumstances, a tax cut's impact on aggregate labor supply would not be indeterminate, but labor inducing, because in the aggregate there is no income effect, only a positive substitution effect.[1]

Gwartney and Stroup illustrate this point by examining the aggregate labor supply effects of an increase in transfer payments financed by an exactly offsetting increase in income taxes. They argue that if the transfer payments are income conditioned, this policy would decrease net wage rates for transfer recipients, as well as for taxpayers, thereby producing work-reducing substitution effects for all members of society. On the other hand, the aggregate change in disposable income in the economy would be zero since, by design, the expansion in transfers is exactly offset by the increase in taxes. They conclude that, unless recipients and taxpayers respond differently to a given change in disposable income, "the income effect of the tax-transfer program will leave both the aggregate supply of labor and consumption of leisure unchanged. [But] Of course, the substitution effect will still be present. Unambiguously, it will induce individuals to work less, although how much less is strictly an empirical issue" (p. 450).

In this comment, we demonstrate that the G-S tax-transfer illustration and the policy implications drawn from it are highly misleading. We first show that the G-S conclusions about the relationship between tax rates and labor supply are based upon strong assumptions about the degree of homogeneity in the population and that without these assumptions, their conclusions are not guaranteed to hold. We then use microsimulation techniques to challenge the impression given by the G-S paper that more redistribution must necessarily imply a reduction in the labor supply and output of the economy.

I. Formalization of the G-S Tax-Transfer Illustration

In order to clarify Gwartney and Stroup's argument concerning the tax-transfer illustration, it is instructive to examine it first within a highly simplified context. We thus begin with a one-consumer economy within which a single consumption good is produced under constant returns to scale. Assuming the individual has only T hours of time available, the production possibility frontier between the consumption good and leisure is represented by the line $TEAB$ in

*University of Notre Dame, Notre Dame, IN 46556, and University of Maryland-Baltimore County, Catonsville, MD 21228, respectively. We are indebted to Gary Burtless, Robert Moffitt, James Rakowski, Philip Robins, and an anonymous referee for helpful and valuable comments on earlier drafts of this paper.

[1]A virtually identical argument is made by Ronald Ehrenberg and Robert Smith (1982, see pp. 151–52). One important difficulty with this argument, which we do not emphasize in this paper, is that there is no particular reason to expect most workers respond similarly to a dollar change in government expenditures on a public good as to a dollar change in personal disposable income. For example, how many workers would increase their labor supply if the government finds itself unable to purchase an aircraft carrier because of a tax cut?

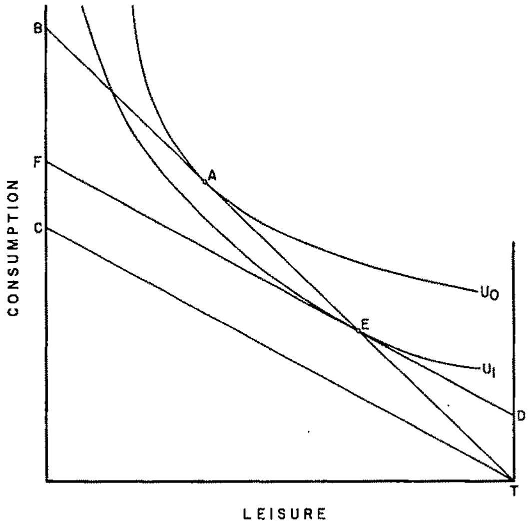


FIGURE 1

Figure 1. In the absence of any government, the individual is in equilibrium at point A and achieves a level of well-being of U_0.

Now consider the implementation of a proportional tax on income. If it is assumed that the individual ignores what is done with the tax revenue, the relevant budget constraint is TC and the individual will work more or less than before, depending upon the relative size of the income and substitution effects resulting from the tax. It is Gwartney and Stroup's contention that the fallacy of this analysis is precisely the assumption that the individual ignores what is done with the tax revenue. A simple way to account for the tax revenue is to assume that the government returns them to the individual. Hence, if the government balances its budget, the economy will be in equilibrium where, due to the convexity of his preferences, the individual works less than he did in the absence of the tax and transfer scheme. In Figure 1, this new equilibrium is depicted by point E on the budget constraint DEF.

Having constructed this simple example to illustrate the G-S argument, we now utilize it to highlight some confusing statements in their article. First, they state (p. 450) that a balanced budget transfer by the government will leave the aggregate disposable income and the total production of society unchanged. However, as our simple example illustrates, if there are reductions in hours of work and in earnings, both the economy's output and disposable income must fall, even though the government has balanced its budget. Aggregate disposable income will remain unchanged only if aggregate gross earnings in the economy remain unchanged. A second and related point of confusion relates to what Gwartney and Stroup refer to as a substitution effect. Traditionally, substitution effects have been defined in terms of the effect of changes in the wage rate holding either utility or disposable income constant. However, as indicated by Figure 1, neither are held constant in the G-S tax-transfer illustration. The substitution effect that Gwartney and Stroup refer to in their article is in reality the effect on labor supply of changing wage rates while holding the production possibilities of the economy constant. This aggregate "Friedman" substitution effect will be negative as long as the set of community indifference curves are convex and the production possibility frontier displays a positive relationship between aggregate hours of work and the production of consumption goods.[2]

In addition to clarifying some confusing statements, our simple example also serves to highlight three assumptions that are essential to the G-S argument, but not made explicit in their discussion. The first of these is that everyone can adjust their hours in response to the implementation of a balanced budget transfer program so as to obtain their desired level of work. One important, if obvious, violation of this assumption occurs when some individuals do not work prior to the policy change. If the nonworkers have a higher probability than do workers of being recipients of transfers, the work-inducing income effects that are engendered by the policy for taxpayers will not, as Gwartney and

[2]We denote this variety of substitution effect as a Friedman substitution effect because to the best of our knowledge it was initially introduced in Milton Friedman's work on the Marshallian demand curve. See his 1949 article and 1976 study for the use of an identical argument that analyzes an excise tax.

Stroup argue, be offset by the work-reducing income effects to which transfer recipients are subject. To see the importance of this, assume hypothetically that there are only two types of individuals in the economy: those who work and those who cannot work due to a disability. Further, assume that the government institutes a transfer program for the disabled population financed by a tax on workers. In this case, the economy will still be on its production frontier. However, it is impossible to determine a priori whether the labor supply among the workers, and hence, the economy as a whole, will increase or decrease.

A second essential (but implicit) assumption to the G-S argument is that the population must be homogeneous with respect to their productive capabilities. This assumption is necessary to guarantee that there exists a positive relationship between the production of consumption goods and aggregate hours of work. However, given existing differences in relative efficiency of workers in our society, it is doubtful that this relationship actually occurs under all circumstances. Consider, for example, an economy with workers of differing productive capabilities. Assume that a tax-financed transfer program has been introduced into this economy and, as a consequence, aggregate hours fall. Are we to infer that production necessarily also falls? Obviously, if each worker reduces their work effort, production must fall. However, it is possible that relatively high-wage workers increased their hours of work, although by less than the amount that low-wage workers decrease their hours. Under these circumstances, it is possible for aggregate production (gross earnings) to rise. Because earnings better reflect the value of output than do hours of work, we feel that earnings are a more natural and policy relevant measure of work effort. Moreover, earnings provide a superior indicator of the effects of labor supply on tax revenues and transfer outlays.

A third essential assumption is that a set of convex community indifference curves exists. This assumption is needed to assure that when the economy moves along the production possibility frontier, the Friedman substitution effect will be negative at the aggregate level. The existence of community indifference curves is dependent upon whether or not the individual preferences in society can be aggregated up to a representative consumer. For an exact linear aggregation to be feasible when individual preferences are specified in terms of consumption of goods and leisure, three conditions must be met. First, all consumers must face the same prices for consumption goods and leisure (wage rates). Second, individual preferences do not have to be entirely homogeneous, but they must all be quasi homothetic; that is, all individual demand functions must be linear with respect to income. Third, when prices are held constant, all individual preferences must display equal and constant marginal utilities of income.[3]

While all three of these conditions are fairly restrictive, the first restriction is obviously violated in the U.S. economy because of differences in wage rates. However, as suggested above, aggregate earnings are a more natural and policy relevant measure of work effort than hours of work, since such a measure is proportional to labor supply measured in efficiency units. Exact linear aggregation over consumption goods and earnings is feasible, even when differences in wage rates exist, if individual preferences are of the form of the Linear Expenditure System.[4] Hence, in order to guarantee that community indifference curves will exist, Gwartney and Stroup must rely upon fairly restrictive assumptions about individual preferences, assumptions that are unlikely to be met in the economy.

The purpose of this section was to outline the conditions and assumptions under which the G-S conclusions concerning their tax-transfer illustration are valid. In summary, these conditions require a fairly homogeneous population with respect to preferences

[3]See Angus Deaton and John Muellbauer (1980, pp. 149–53) for an extended discussion of the conditions required for an exact linear aggregation of individual preferences to be feasible.

[4]See Deaton and Muellbauer (pp. 159–60) for a discussion of exact linear aggregation of individual preferences with endogenous labor supply.

and productive capabilities. It is our opinion that these conditions are so restrictive and unrealistic that at best the G-S conclusions are merely of theoretical interest. Stated slightly differently, once heterogeneity in the population is allowed to enter the analysis, the direction, as well as the magnitude, of the labor supply effects of balanced budget changes in government policies must be viewed as an empirical question.[5]

II. An Illustrative Simulation

In this section, we present an illustrative microsimulation of a hypothetical redistribution program that is fully financed through taxes. As will be seen, results from this simulation suggest that under certain plausible circumstances, schemes that redistribute income to the low-income population can produce increases in aggregate labor supply. We feel this result, although tentative, is important because Gwartney and Stroup, as well as others (see Edgar Browning and William Johnson, 1984, for example), have tended to leave the impression that such schemes must inevitably reduce aggregate hours of work. Our simulation results are intended as a counterexample to this view.

In conducting the simulation, we have tried to be as realistic as feasible. For example, the simulation does not allow individuals to reduce their hours of work to less than zero, even though they may "desire" negative hours of work. The simulation also permits male and female labor supply functions to differ. For purposes of consistency with the G-S analysis, however, the simulation is based upon the assumption that both male and female labor supply functions are linear with respect to income.[6]

The hypothetical income redistribution scheme we simulated would replace both the federal income tax and most existing federal welfare programs (including Aid to Families with Dependent Children, Supplemental Security Income, and Food Stamps) with an integrated and simplified tax-transfer system. This plan would guarantee a minimum income of \$2,550 per adult and \$1,275 per child (in 1984 dollars), an amount that would provide a family of four an income guarantee equal to 75 percent of the poverty line.[7] The transfer payments would be income conditioned, reducing the payment by 50 cents for every dollar of income the household receives from other sources. As a consequence, marginal tax rates faced by most, but not by all, of those who receive transfers under this plan would increase.

Since total transfer payments under this plan would be larger than those currently made, the tax system would have to collect additional revenues. The implemented plan would do this in four ways. First, unlike the current tax system, all Unemployment Compensation and Social Security payments would be included in the definition of gross income. Second, all deductions from adjusted gross income would be eliminated. Third, personal exemptions would be raised to \$5,100 per adult and \$2,550 per child. Finally, taxable income would be taxed at a flat 23 percent rate, a rate just sufficient to raise enough revenues to leave the government's budget unchanged when compared to prior to the reform. This tax rate represents

[5]The conditions we have listed above should be viewed as sufficient for the G-S conclusion to hold. They do not represent necessary conditions. For example, a class of preferences that cannot be aggregated up to a representative household, yet can still be shown to yield the G-S conclusions when wage rates vary across households, is the parallel utility functions (see Jonathan Dickenson, 1979, or William Gorman, 1961). This class of preferences possesses parallel income consumption curves.

[6]The labor supply functions utilized in the simulation were estimated from data collected in the Seattle-Denver Income Maintenance Experiments. These parameter estimates have previously been used to conduct several transfer program simulations reported in this *Review* (see Michael Keeley et al., 1978, and Gary Burtless and Greenberg, 1982). For a more detailed description of the simulation model and procedures, see our paper with Richard Kasten (1980).

[7]Aged, blind, and disabled adults are eligible for \$3,315 in terms of income guarantee and are given a personal exemption of \$6,630 in the tax system.

an increase in the marginal tax rate faced by most, but not all, nontransfer recipients under the plan.

The simulation of this hypothetical tax-transfer system suggests that its adoption would result in considerable redistribution of income. For example, prior to any labor supply response, the Gini coefficient would fall by over 10 percent (from .456 to .406) and the number of persons in poverty would fall by over 25 percent. But the simulation also predicts that this redistribution would engender an increase in labor supply, with total hours in the nation rising by 1.1 percent and total earnings by 3.5 percent. These findings, which are generally supported by additional simulations we have conducted, strongly suggest that tax-financed transfer programs will not necessarily decrease aggregate labor supply.[8]

III. Conclusions

The purpose of this note has been twofold. First, we wished to demonstrate that Gwartney and Stroup's conclusion that aggregate labor supply *must* fall if increased transfer payments are financed through increased taxes is based upon fairly strong assumptions about the homogeneity of preferences and productive capabilities in the economy. It is our opinion that if realistic assumptions about heterogeneity in the population are made, predictions about the direction of the aggregate labor supply effects of such policies cannot reliably be made solely upon the basis of economic theory, as G-S suggest, but must be resolved empirically. A second and related objective of the comment was to utilize microsimulation techniques to illustrate the possibility of a counterexample to the implication in the G-S article that increased transfers will lead to reductions in aggregate labor supply. Although findings from these simulations should be considered tentative—indeed, the general equilibrium effects of income redistribution policy is a potentially fruitful area for future research —they are quite suggestive. They imply that tax-financed increases in transfer payments are unlikely to have a strong adverse effects on hours worked and earnings in the economy; indeed, positive effects appear feasible.

[8] The results presented in the text are from only one set of labor supply parameter estimates and one tax-transfer scheme. To examine the robustness of these findings, we conducted a number of additional simulations based on different labor supply parameters and alternative tax-transfer schemes. Several of these additional simulations are reported in our paper with Kasten (1982), and others are available from us upon request. In general, these simulations indicate that income redistribution would be associated with higher aggregate hours and earnings, although the predicted magnitudes vary considerably. And in those relatively rare instances when hours or earnings were predicted to decline, the estimated reductions were quite small—always under 1 percent. A very conservative conclusion from these simulations is that reductions in labor supply need not be a major obstacle to substantial redistributions of income.

REFERENCES

Betson, David, Greenberg, David and Kasten, Richard, "A Microsimulation Model for Analyzing Alternative Welfare Reform Proposals: An Application to the Program for Better Jobs and Income," in Robert Haveman and Kevin Hollenbeck, eds., *Microeconomic Simulation Models for Public Policy Analysis*, Vol. 1, New York: Academic Press, 1980, ch. 5.

______, ______, **and** ______, "A Simulation Analysis of Economic Efficiency and Distributional Effects of Alternative Program Structures: The Negative Income Tax Versus the Credit Income Tax," in Irwin Garfinkel, ed., *Income-Tested Transfer Programs: The Case For and Against*, New York: Academic Press, 1982, ch. 6.

Browning, Edgar and Johnson, William, "The Tradeoff between Equality and Efficiency," *Journal of Political Economy*, April 1984, *92*, 175–203.

Burtless, Gary and Greenberg, David, "Inferences Concerning Labor Supply Behavior Based on Limited-Duration Experiments," *American Economic Review*, June 1982, *72*, 488–97.

Deaton, Angus and Muellbauer, John, *Economics*

and Consumer Behavior, Cambridge: Cambridge University Press, 1980.

Dickenson, Jonathan, "Revealed Preference, Functional Form and Labor Supply," Working Paper 30-78, Department of Economics, Pennsylvania State University, 1979.

Ehrenberg, Ronald and Smith, Robert, *Modern Labor Economics: Theory and Public Policy*, Glenview: Scott, Foresman, 1982.

Freidman, Milton, "The Marshallian Demand Curve," *Journal of Political Economy*, December 1949, *57*, 463–95.

______, *Price Theory*, Chicago: Aldine Press, 1976.

Gorman, William, "On a Class of Preference Fields," *Metroeconomica*, August 1961, *13*, 53–56.

Gwartney, James and Stroup, Richard, "Labor Supply and Tax Rates: A Correction of the Record," *American Economic Review*, June 1983, *73*, 446–56.

Keeley, Michael et al., "Estimation of Labor Supply Models Using Experimental Data," *American Economic Review*, December 1978, *68*, 873–87.

Labor Supply and Tax Rates: Reply

By JAMES D. GWARTNEY AND RICHARD L. STROUP*

David Betson and David Greenberg (B-G) add some interesting refinements to what we have said in our 1983 paper. In doing so, they incorporate in their analysis our major point: it is inappropriate to apply the income effect of the traditional analysis, as several prominent economists whom we quoted did, without also considering the income effect of offsetting changes in government expenditures. Our earlier analysis assumed that equal and opposite changes in income for taxpayers and recipients will exert equal and opposite effects on aggregate earnings. Betson and Greenberg correctly introduce the offsetting income changes from increased tax transfers. Then, in contrast to us, they argue that the positive earnings effects of the taxpayers' reductions in income exceed the negative labor supply effects of the recipients' increase in income. Their simulation results also suggest that with substitution effects reduced by a flat tax rate of 23 percent and a much lower implicit tax rate on transfer recipients than typically exists now for those receiving large transfers, increased transfers might actually expand aggregate earnings.

Specifically, Betson and Greenberg assume that recipients of enlarged transfer programs respond to their income increases by reducing earnings by less than taxpayers increase theirs in response to the corresponding income reductions. These responses are plausible, especially within their simulated world of a low, flat tax rate and a negative income tax program. Their program works to minimize the large substitution effect caused by high tax rates and high implicit tax rates on the poor that now characterize the real policy world. So when they also specify a strong income effect for taxpayers, and a weak income effect for transfer recipients, the total result is unsurprising. While their theoretical point is correct, its policy relevance is another question.

The B-G special case of transfers only to disabled individuals, whose production is zero, also has limited relevance. First, at least four of every five federal transfer dollars are in programs not even means tested, much less disability tested. Second, to the vast majority of voters who must approve them, transfers to the disabled seem logically to be in themselves a good. So an increase in them at the margin may politically indicate that voters (most of whom are taxpayers) are not made worse off by the larger program. Once again, there need not be any income effect at all, even for taxpayers per se. Third, the taxpayer who does not want more of his dollars going from the government to the poor may well have another option: reducing his private donations to the same cause. Russell Roberts (1984) presents evidence indicating that historically, when transfer programs for the poor have expanded, private charity given to the poor has contracted. Tax-transfer programs are a substitute for private charity, and may crowd out that charity. Again, if taxpayers do not feel poorer, there may be no income effect from taxes to provide aid to the nonproductive. But the substitution effect remains, and is unambiguous in its direction. Fourth, even transfers to the disabled go to households which may have productive members, so that an offsetting income effect is possible.

Still, Betson and Greenberg are correct to say that their changes in our assumptions do cause the net effect on total output of greater tax-transfer programs to become an empirical question. However their amendment of our analysis does not negate our major point—that the income effect from government expenditures tend to offset the income effect from taxes to finance those expenditures.

*Departments of Economics, Florida State University, Tallahassee, FL 32306-2018, and Montana State University, Bozeman, MT 59715, respectively.

REFERENCES

Betson, David M. and Greenberg, David, "Labor Supply and Tax Rates: Comment," *American Economic Review*, June 1986, *76*, 551–56.

Gwartney, James D. and Stroup, Richard L., "Labor Supply and Tax Rates: A Correction of the Record," *American Economic Review*, June 1983, *73*, 446–56.

Roberts, Russell, "A Positive Model of Private Charity and Public Transfers," *Journal of Political Economy*, February 1984, *92*, 136–48.

The Disinterest in Deregulation: Comment

By MARTIN CHERKES, JOSEPH FRIEDMAN, AND AVIA SPIVAK*

In a recent article in this *Review*, Robert McCormick, William Shughart, and Robert Tollison (1984), following the seminal contributions of Arnold Harberger (1954) and Gordon Tullock (1967), addressed the issue of monopoly deregulation. The authors put forward the provocative idea that "because under most conditions Tullock costs cannot be recouped, the returns to deregulation are lower than previously thought" (p. 1075). In particular, they claim that because most rent-seeking expenditures are made before monopolies are established and are largely *sunk costs*, little is gained from dismantling monopolies.

In this comment we challenge both assumptions. We point out that McCormick et al. have not actually shown, as they claim, that "such a deregulating program can easily impose more costs than it is worth" (p. 1075).

Section I examines the nature of Tullock costs and their distribution throughout the life of a monopoly. We conclude that only a portion of these costs may be considered sunk. Section II provides a short case study of the American Medical Association's (AMA) ascent into monopolistic power. While obtaining the monopoly did not require substantial resources, we show that the ongoing maintenance of the monopoly involves large rent-seeking costs. In Section III we correct an error in McCormick et al.'s diagrammatic exposition and surmise that even if all Tullock costs are sunk, deregulation may be still desirable because it redistributes income from monopolies to consumers.

*Cherkes: Department of Finance, Temple University; Friedman: Department of Economics, University of Pennsylvania, Philadelphia, PA 19104; Spivak: Department of Economics, Ben Gurion University.

I. Are Most of the Tullock Costs Sunk?

According to Tullock, monopolists spend resources to create and maintain their monopoly power. Richard Posner (1975) added structure to Tullock's idea by assuming that obtaining the monopoly is a competitive activity, with constant unit cost and no socially valuable by-products. Thus, an amount equal to the discounted value of the monopoly rent will be spent on obtaining and maintaining the monopoly privilege. Posner named this cost "Tullock cost," T. Together with the familiar Harberger triangle of the loss of consumer's surplus, H, the total cost of monopoly is $T+H$.

Tullock did not address the allocation of the Tullock cost over time, but McCormick et al. assume that most of the cost is incurred while the monopoly is formed and that current, maintenance, cost is negligible.

The allocation of costs over time deserves more attention than McCormick et al. or Posner gave it. We will show why a substantial fraction of the Tullock cost is incurred after the monopoly is established. We distinguish between two groups of individuals who spend resources to get access to the monopoly privilege. The first group is the monopoly "founders." Founders spend resources to create the monopoly and later contribute to its maintenance. The second group are late "joiners" to the monopoly. Both founders and joiners allocate resources for rent seeking. Resources are allocated as any other resource intended to increase profit (for example, advertising). That is, in deciding how much resources to allocate for rent seeking, a representative founder considers the potential benefits and costs. Benefits are the increase in the present expected discounted value of profit and arise by limiting the industry's output below the level that a competitive industry would produce. In

calculating potential profits, founders anticipate the eventual appearance of joiners. This possibility is reflected in the benefit function of founders. When an industry faces downward-sloping demand for its product, both average and marginal benefit functions are downward sloping. Additionally, the area under the marginal benefit function equals the above-normal profits that will arise from rent seeking. The level of rent-seeking activity is determined by the intersection of the marginal benefit and marginal cost functions. The area between the two curves represents the founders' surplus. The surplus is necessarily positive (otherwise the founders would not have created the monopoly) and can be reclaimed by consumers if the monopoly is dismantled.

The surplus exists because of information asymmetry. In the short run, insiders or current practitioners enjoy economic and legal advantages. During the organizational period, outsiders are not fully aware of the emerging monopoly, and even if they are, they cannot join the industry with short notice. However, once the monopoly has been established, everybody is well informed and no one enjoys an advantage. Competition for the surplus will eliminate any profits. This is the case Posner discussed, in which all the profits of the monopoly turn into rent-seeking costs. We suggest that only the expenses incurred by founders should be considered a sunk cost, while the joiners' expenses should be considered an ongoing cost that may be recovered by deregulation.

Our conclusion is strengthened by considering the ways many monopolies are formed. A notable example is OPEC which was formed in 1960 with limited goals and little set-up costs. It became a price-setting cartel only in 1973, exploiting the geopolitical developments in the Middle East. Other examples are the drug industry, where regulation was imposed against the industry's opposition and the regulation of civil aviation in the early 1930's which was initiated to ensure "proper" airmail distribution and air-transport service.[1]

At its inception, the supply of the commodity sold by the monopoly is difficult to reduce, hence monopoly profits are low. It is only later that the supply is reduced and above-normal profits materialize. This pattern exists because supply is often reduced by limiting entrance in face of growing demand. In other cases, the move to limit quantities may come years after the formal organizational framework was established. In fact, governments sometimes create organizations or regulate industries for their own interests without any lobbying from the industry. Eventually, output is limited and profits soar.

To summarize the discussion: monopolies incur significant current rent-seeking costs, often larger than the set-up sunk costs. Moreover, for the founders the cost of the rent-seeking activity is smaller than the expected profit, because entry is not free. Finally, all the rent-seeking activity of those who try to join the monopoly is not included in the founders' sunk cost. The cost of rent-seeking activity *is* equal to the joiners' profit because entry to the activity is free.

II. The Licensing of Physicians

According to Paul Starr, the AMA, which was founded in 1846, had very little impact during the first half century of existence, had scant resources, and failed in its effort of voluntary reform of medical schools: "The irregular physicians accused it of attempting to monopolize medical practice... the irregular thrived" (1982, p. 91). Licensing of physicians was restored only in the 1870's and 1880's.[2]

Licensing was part of a general licensing movement that included plumbers, barbers, horseshoers, etc., and did not arise from the political power of physicians (Starr, pp. 102–03).[3] Initially, licensing did not limit the supply of practitioners, because under the statutes anyone who practiced medicine

[1]See Henry Kissinger (1982) and Albert Danielsen (1982) on the history of OPEC, Tullock (1975) on the drug industry, and Richard Caves (1962) on the airline industry.

[2]See George Stigler (1971) for details.

[3]Earlier attempts during the 1830's and 1840's by physicians to require licenses were unsuccessful (Starr, p. 62).

at the time was qualified. In fact, the increased demand for medical degrees created a supply of new medical colleges.

The account of the development of licensing indicates a pattern which is different from the one suggested by McCormick et al. The profession did not unite and did not spend significant resources to obtain the licensing privilege. The privilege fell in their laps when the time was ripe with only negligible sunk costs. The ascent of the medical profession came later, due to major scientific developments in the late nineteenth century (vaccination, sterilization, etc.). The medical profession underwent a period of self-organization contemporaneously with other movements, such as labor. Medical schools were reformed as part and parcel of the reforms of universities starting in 1870; more stringent standards were applied to medical degrees only after the Flexner report of 1906 (Starr, pp. 112–30). The monopoly was established only after the supply of physicians fell behind population growth (Starr, pp. 126–27). This new monopoly power was reflected in the rapid increase in the physician's income in the early twentieth century (Starr, pp. 142–43).[4]

When the activities which gave rise to the physicians' monopoly are examined, political lobbying does not appear high on the list. What we see is a profession reorganizing its schools and its association, activity which contained substantial side benefits. The rent-seeking costs are not in lobbying because the medical profession has established itself as an indisputable authority on health care. Rent-seeking costs associated with the medical monopoly are of two kinds: the first are the resources wasted on competition to join the medical profession. The resources include the efforts of those who attempt to be admitted to medical schools and fail, in excess of the level that would prevail in competition, and the excessive time and effort in medical schools, residencies, etc. If potential physicians have excessive discounted expected return on their investment, the marginal candidate to medical school should spend resources whose expected cost is equal to the expected return.[5] The second kind of rent-seeking cost is related to the organization of medical care and to the fact that it may be responsible for the high cost of care, to the insufficient number of doctors in rural areas, to the way the medical insurance is structured, etc. These two types of costs are not sunk costs. They are current costs that may be recouped if the profession was deregulated. Thus, the concept of sunk costs fails to explain the unchallenged privileged status of physicians.

There is an economic-legal rationale for the pattern of regulated monopolies discussed in Section I, exemplified by the medical profession. A regulated monopoly is a group of economic agents in an industry that can limit entrance. In a democratic society, the number of agents who are already in such industry cannot be reduced legally by passing a law. Those who already practice a trade have certain legal rights which often ensure their staying in business. From the legal viewpoint it is much easier (and hence more economical) to limit entrance of *new* agents. It follows that because the immediate gains of forming a monopoly are small and because the future gains are discounted and *uncertain*, founders should be willing to spend little on establishing the monopoly. Later on, when the limitations on entry bear monopoly profit, the founders and prospective entrants should be willing to spend the rent-seeking costs.

III. Even if the Tullock Costs are Sunk, Deregulation is Still Desirable

We have shown that for the physicians' licensing example mentioned by McCormick et al., most rent-seeking costs are not sunk. Since we explained the economic rationale for this pattern in Section I, other examples may show the same temporal structure if analyzed carefully. In this section we argue that even for industries where the premise of sunk cost is acceptable, McCormick et al. have failed to prove their point. Their analy-

[4] See also Milton Friedman and Simon Kuznets (1945, p. 464).

[5] This assumes risk neutrality. A risk-adverse individual will spend less.

sis contains errors which led them to state that a deregulation program can easily impose more costs than it is worth. They have also completely ignored the issue of income redistribution.

A. *Correcting McCormick et al.'s Diagrammatic Exposition*

McCormick et al. assert that the rent-seeking activity of the regulated industry permanently raised the long-run marginal cost schedule from *LRMC* to *LRMC'*, because the rent-seeking activity uses the best resources of the industry (see our Figure 1). Thus, society's gains from deregulation would not be the Harberger triangle *ABC*, but only the triangle *A'B'C*. The trapezoid *ABB'A'* is a sunk cost, lost forever to the economy. In this spirit, if *LRMC* goes above point *C*, the Harberger triangle will disappear. McCormick et al. fail to note that as the *LRMC* curve moves up, the monopoly is adjusting price and quantity and the Harberger triangle moves up as well. The deadweight loss to be regained is *FB'C'* which contains the area *A'B'C*. While the new Harberger triangle *FB'C'* is smaller than *ABC*, it does still exist. The assumption that *LRMC* increases because of sunk cost is rather arbitrary. In many cases, regulation results in excess capacity which lowers the *LRMC*.[6] If this is the case, the Harberger triangle, in fact, will *increase* in area.

B. *Redistribution of Income*

Deregulation of a monopoly redistributes income from the monopoly owners to consumers. The order of magnitudes involved is rather significant. Posner estimated that monopoly profits are about 3.3 percent of *GNP*. In comparison, the federal government, using the Social Security program, redistributes 4.7 percent of the *GNP*.[7] Reducing physicians' income through deregulation would certainly redistribute income in the right direction. From the view of positive economics, there is thus an incentive to consumers to act politically to dismantle a monopoly, even if rent-seeking costs were sunk. Consumers who are furious with their medical bills, or employers who bear large medical insurance costs, have a good incentive to organize politically against the monopoly, irregardless of whether the social costs are sunk. This positive argument is not mitigated by the possibility that doctors have spent excessive resources to obtain their licenses, as pointed out by Tullock (1975).

[6] Harvey Averch and Leland Johnson (1962) is the locus classicus.

[7] See Michael Boskin, Marcy Averin, and Kenneth Cone (1983).

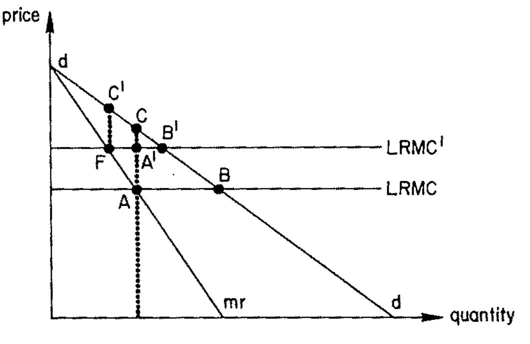


FIGURE 1

REFERENCES

Averch, Harvey and Johnson, Leland, "Behavior of the Firm under Regulatory Constraint," *American Economic Review*, December 1962, *52*, 1052–69.

Boskin, Michael J., Averin, Marcy and Cone, Kenneth, "Modeling Alternative Solutions to the Long-Run Social Security Funding Problem," in Martin Feldstein, ed., *Behavioral Simulation Methods in Tax Policy Analysis*, Chicago: University of Chicago Press, 1983.

Caves, Richard, *Air Transport and Its Regulators*, Cambridge: Harvard University Press, 1962.

Danielsen, Albert L., *The Evolution of OPEC*, New York: Harcourt, Brace and Jovanovich, 1982.

Friedman, Milton and Kuznets, Simon S., *Income from Independent Professional Practice*, NBER General Series, No. 45, University Microfilms, 1945.

Harberger, Arnold, "Monopoly and Resource Allocation," *American Economic Review Proceedings*, May 1954, *74*, 77–87.

Kissinger, Henry, *Years of Upheaval*, Boston: Little, Brown, 1982.

McCormick, Robert E., Shughart, William F. and Tollison, Robert D., "The Disinterest in Deregulation," *American Economic Review*, December 1984, *74*, 1075–79.

Posner, Richard A., "The Social Costs of Monopoly and Regulation," *Journal of Political Economy*, August 1975, *83*, 807–27.

Starr, Paul, *The Social Transformation of American Medicine*, New York: Basic Books, 1982.

Stigler, George J., "The Theory of Economic Regulation," *Bell Journal of Economics*, Spring 1971, *2*, 3–21.

Tullock, Gordon, "The Welfare Costs of Tariffs, Monopolies and Theft," *Western Economic Journal*, June 1967, *5*, 224–32.

______, "The Traditional Gains Trap," *Bell Journal of Economics*, Autumn 1975, *6*, 671–78.

The Disinterest in Deregulation: Reply

By Robert E. McCormick, William F. Shughart II, and Robert D. Tollison*

In our 1984 article, we combined the traditional notion of the welfare cost of monopoly (Arnold Harberger, 1954) with the possibility of an associated rent-seeking cost (Gordon Tullock, 1967) to make a simple point: To the extent that monopoly returns are dissipated by expenditures to obtain a monopoly right, the costs of regulation are sunk, and cannot be recouped *ex post* through deregulation. On the other hand, any ongoing expenditures necessary to maintain or to defend a monopoly, the existence of which signify that the full rent-seeking costs have not been completely borne, are potentially recoverable. Our analysis suggested that the gains from deregulation may be smaller than typically perceived, and that the returns to preventing monopoly in the first place are relatively large. In effect, we offered an alternative explanation for consumer/voter apathy toward economic regulation.

In their comment, Martin Cherkes, Joseph Friedman, and Avia Spivak (1986), argue that we have failed to prove our point. In so doing, they raise a minor issue that we freely admit. Our diagramatic exposition (Figure 3) ignored the output and price adjustments along the monopolist's marginal revenue schedule. But, this is a trifle. Our point that the new Harberger triangle is smaller after regulation *still* goes through.

More importantly, Cherkes et al. use a series of *ad hoc* examples, including the licensing of physicians and the OPEC cartel, to suggest that in the early stages of regulation those lobbying for a monopoly right (the "founders") discount its value heavily because of low initial profits, uncertainty about the future, and the realization that new members will eventually have to be admitted to the industry (the "joiners"). These examples, which are by no means definitive histories of these interest groups, are just an illustration of one implication of our analysis. As we said earlier, if a large portion of the costs of regulation are ongoing, there is more to be gained from deregulation than any standard analysis in the literature predicts and vice versa. The reader should not be confused; this is one of our basic points, not theirs.

Cherkes et al. assert, however, that even if all rent-seeking costs are sunk, deregulation is still desirable because it transfers income to consumers. There are two cases to consider. First, suppose that the discounted value of the monopoly has been capitalized by, for example, its transfer from founders to joiners, who subsequently earn a normal rate of return. In this case, a deregulatory program that redistributes income to consumers must do so by confiscating a portion of the joiners' wealth. Although society would on net gain the value of the adjusted Harberger triangle from such a deregulation, the overall welfare implications are far from clear. The redistribution at stake here is not a redistribution from the original monopolist to consumers; it is from the "little old ladies" who bought stock in the cartel after it was formed to consumers in the mass. Second, if capitalization has not taken place, the gains from deregulation amount to the value of the adjusted Harberger triangle plus that portion of the rectangle of monopoly profits not yet dissipated by rent-seeking activities. In short, it is always true that the private benefits to consumers from deregulation can exceed their private costs, but this has really nothing to do with the main point of our analysis. Also, Cherkes et al. seem confused about opportunity cost here. If, as they argue, consumers "...furious with their medical bills..." are

*McCormick: Department of Economics, Clemson University, Clemson, SC 29631; Shughart and Tollison: Center for Study of Public Choice, George Mason University, Fairfax, VA 22030.

politically potent enough to lobby successfully for a rent redistribution in their favor, why are they not politically powerful enough when they face a weaker opponent, a competitive medical industry? Such considerations illustrate—as we emphasized—that the returns to preventing regulation in the first place are relatively large.

In sum, nothing in the comment by Cherkes et al. moves us to alter our original analysis. The distinction between the costs of monopoly and the costs of monopolization is now well established in the literature, and the positive theme of our paper provides an alternative explanation for the apparent lack of interest in deregulation. If, as Cherkes et al. would have it, there is much to be gained from reducing the amount of the monopoly in the world, why are there so few empirical examples of deregulation in comparison with the long and quiet history of regulation?

REFERENCES

Cherkes, Martin, Friedman, Joseph and Spivak, Avia, "The Disinterest in Deregulation: Comment," *American Economic Review*, June 1986, *76*, 559–63.

Harberger, Arnold, "Monopoly and Resource Allocation," *American Economic Review Proceedings*, May 1954, *44*, 77–87.

McCormick, Robert E., Shughart, William F. II and Tollison, Robert D., "The Disinterest in Deregulation," *American Economic Review*, December 1984, *74*, 1075–79.

Tullock, Gordon, "The Welfare Costs of Tariffs, Monopolies, and Theft," *Western Economic Journal*, June 1967, *5*, 224–32.

Internal Migration and Urban Employment: Comment

By MICHAEL P. TODARO*

The recent paper by William Cole and Richard Sanders (1985) on migration and urban employment in developing countries is seriously flawed by a number of statistical and conceptual inaccuracies, extreme and arbitrary assumptions and, *ad hoc*, unsubstantiated policy conclusions. Since the article is largely devoted to a critique and proposed modification of my migration model (1969), a comment and response seems appropriate.

The authors begin by asserting that "this paper...presents data amassed *from a wide range of countries* to highlight a crucial shortcoming of that [Todaro] model" (p. 481, emphasis added). Such a statement normally leads one to expect lengthy statistical tables, copious literature references, and diverse empirical studies as part of that "amassed" data. Instead, what we get are a few sporadic references to outdated, secondary Mexican statistics, no statistical tables, no summary of empirical research (in fact, on p. 494, Cole-Sanders, in their only reference to empirical tests of the Todaro model, find that "the model has been found consistent with the data"), and no data on internal migration and urban employment from any of the 143 nations of the Third World. Cole and Sanders then claim that they will also show that "far from being general in nature, the Todaro approach is limited to explaining the movement of persons possessed of sufficient human capital to qualify them for modern sector employment" (p. 481). They achieve this demonstration by the arbitrary nature of the way *they* define the urban modern (*U-M*) and urban subsistence (*U-S*) sectors, effectively excluding *by definition* anyone without "human capital" from the *U-M* sector even though they may earn *U-M* wages. While a number of well-known modifications of the original Todaro and Harris-Todaro models have been published over the past sixteen years—including outstanding theoretical papers by Gary Fields (1975), Joseph Stiglitz (1974; 1976), Jagdish Bhagwati and T. N. Srinivasan (1974), W. Max Corden and Ronald Findlay (1975), George Johnson (1971), Peter Neary (1981), M. Ali Khan (1980), Thomas McCool (1982), among many—it is difficult to recognize the theoretical significance of the Cole-Sanders paper (there being no attempt at empirical or methodological contribution). In fact their idea of segmenting the urban informal (*U-S*) from the formal (*U-M*) sector and distinguishing between workers with and without human capital has been successfully and more convincingly dealt with algebraically in the literature as far back as the 1975 Fields paper (for which no reference is even given in the Cole-Sanders paper), as well as indirectly in papers by Richard Sabot (1975; 1977), W. F. Steel and Y. Takagi (1976), Dipak Mazumdar (1976a, b), myself (1972, 1976), and others. However, in contrast to Cole-Sanders, previous authors recognized the interdependence of the locational choice of urban migrants, that is, some go directly to the modern sector, others get jobs in the informal sector while engaging in modern sector job search, while still others come to the city hoping for modern sector jobs but remain permanently in the unorganized informal sector. The Cole-Sanders' theory requires them to create an artificial separation in which some migrants go only to the *U-M* sector, while others go only to and never leave the *U-S* sector.

Finally, the authors claim to "develop a model that serves as a useful complement to that of Todaro" (p. 481). It turns out that there really is no new "model" to speak of in the sense of an original set of internally consistent and realistic equations yielding unique algebraic solutions that shed new light

*Department of Economics, New York University, New York, NY 10003, and Center for Policy Studies, The Population Council, New York.

on an otherwise contentious economic problem. There is only a series of numbered relationships based on extremely restrictive assumptions like perfectly elastic *U-S* labor supplies, perfectly inelastic *R-S* labor supplies, no *U-S* or *R-S* savings, fixed agricultural prices, instantaneous employment of all *U-S* migrants, and an equilibrium condition that goes to great lengths to make the obvious point that in a competitive world with free factor mobility and full employment, wages will be equalized between *U-S* and *R-S*!

I will deal first with the authors' criticism of my model and then discuss their proposed modifications. Cole and Sanders' basic criticism of my model seems to be that it yields implausible time horizons for eventual equality in expected urban modern sector and average rural incomes, thus vitiating its usefulness in explaining why a potential migrant with little education decides to take his chances in the urban modern labor market. They attempt to demonstrate this by means of Tables 1 and 2 in which they engage in a numerical exercise, using what they claim to be "plausible" but hypothetical values for some of the key parameters of my basic model. Aside from the fact that the original model was not designed to provide a literal prediction of the number of years needed to bring about equality in expected incomes (a number of earlier commentators dealt with that issue), but rather to demonstrate how in the absence of wage flexibility the urban unemployment rate would ultimately act as the equilibrating factor in regulating migration and to show why rapid migration continues in the face of high and rising urban unemployment, the model's predictive accuracy is not as poor, even in its literal numerical interpretation, as Cole and Sanders make out in Table 2. The fact is that they have greatly exaggerated the results in Table 2 by using an incorrect statistic for the urban modern-rural subsistence income differential, and by failing to use a discount rate that is the appropriate one for typical *LDC* urban migrants. Even using the figures for π in the .12 to .18 range of Table 2, if Cole and Sanders had researched the actual magnitude of the urban-rural income gap, they would have discovered that the gap is not 2-to-1 as they assume in the table but more like 4 or 5-to-1. For example, Michael Lipton (1977, p. 435–37) has shown that in 40 out of 44 *LDC*s in Africa and Asia, the ratio of urban to rural average incomes is *greater* than 2.7, and that in 36 out of 63 *LDC*s (more than half) throughout the world, the ratio was *above* 4-to-1. Moreover, if only urban modern sector incomes were compared to average rural incomes, the disparity would be even greater. In fact, Cole and Sanders themselves inadvertently reveal their presumption of the large and widening urban-rural income gap without incorporating it into their numerical exercise when in Table 3 they characterize average income in the urban modern sector to be "high and growing" while that of the rural subsistence sector is "low and stagnant." Also, rather than using a discount rate of 5, 10, or 15 percent as in Table 2, the authors should have realized that in *LDC*s, the effective discount rate is closer to zero and real interest rates are in many cases negative due to rapid inflation and administered nominal rates. The net effect of either of these two changes (but especially the large and growing wage gap) along with the fact that the real size of the urban surplus labor pool is of the order of .20 to .30 in most *LDC*s means that, contrary to the Cole-Sanders' assertion of 50 or more year time horizons, a more realistic calculation would yield at worst time horizons of two to six years. While these figures better fit the actual work experiences of most unskilled urban migrants who eventually find modern sector jobs as revealed by numerous migrant surveys, I repeat that I never claimed that as simple and aggregative a model as mine should provide a precise numerical prediction of either equilibrium expected income time profiles or equilibrium unemployment/underemployment rates (my 1976 book, p. 38). The Cole-Sanders numerical exercise is thus highly contrived and artificial.

Let me turn now to Cole-Sanders' own framework as presented in Sections IV through VI. They begin by asserting that rural-urban migration is a "dual phenomenon" with educated migrants bound for the

modern sector (Cole-Sanders never say *how much* education) and uneducated ones aspiring only to urban subsistence sector jobs. Is this supposed to be a statement of fact or an assumption? If it is a statement of fact, where is the evidence from the scores of empirical migration studies carried out over the past fifteen years to support this assertion? None is given. My reading of that literature indicates that while some migrants come with no aspirations beyond the subsistence sector, most hope to someday get a high paying modern sector job (see Fields, 1986). So the dual migration phenomenon is in fact an assumption—and a highly unrealistic one at that—by the authors. Yet the whole thrust of the Cole-Sanders argument hinges on this assumption of a strict dichotomy between migrants with human capital and those without. For their argument to make any sense at all, they are forced to assume that the pool of labor from which the flows given by equations (2) and (3) are drawn are completely distinct. Casual references to the fact that a short-run model should not be concerned with the decision to invest in human capital are an inadequate justification for this extreme dichotomy. The formal model thus has little value as a tool of analysis and/or policy.

A few words also need to be said about the existence of "perverse markets" in Third World countries—a phenomenon that seems to deeply disturb the authors (p. 492). The fact is that a great deal of the literature of development economics deals with this perversity and the resulting dichotomy between private and social benefit-cost calculations. Whether in the realm of the economics of family fertility, the costs and returns of investing in primary vs. higher education, the impact of factor price distortions on resource allocation or the ranking of public and private investment projects, the messy and complex world of development economics is permeated by such "perversities." Their existence requires flexibility in our theoretical modeling and policy advice. And, as in the case of my migration model, it requires the development theorist to work within the institutional realities of Third World markets and nonmarkets (in this case, to recognize the downward stickiness of urban modern wages) and analyze alternative mechanisms of economic adjustment (for example, by reference to an equilibrium unemployment rate rather than an equilbrium wage rate as in the unrealistic flexible wage model).

Finally, in the last section of their paper, Cole and Sanders grant that my model "has been found consistent with the data" and "has yielded the correct prediction" with regard to the growth of the *U-S* sector labor force. Despite this recognition, they declare the theory inappropriate and unable to provide a satisfactory explanation for the rapid growth of Third World cities. They assert that the rapid urbanization of the Third World is both privately and socially desirable, despite the existence of a large body of empirical research to the contrary, as well as a series of United Nations surveys of population issues in the Third World showing that most governments believe that their population distributions are "highly unacceptable" with over 80 percent saying they were trying to slow down or reverse their rural-urban migration (United Nations, 1978, pp. 27–28). Even in Mexico, the country used by Cole and Sanders to support their arguments, the government in its 1984–88 National Development Program has asserted that "the concentration of population and economic activity in the Mexico City metropolitan area has reached extremes that distort the social and economic development of the nation" (*The New York Times*, September 8, 1985, p. 26E). The Mexican government blamed lack of rural development, urban bias, and excessive migration as the major reasons for these distortions.

Given the need for simplifying assumptions, we cannot place unreasonable demands on simple or, indeed, most complex models. But, as always, evaluations of theories must focus on the reasonableness of their underlying structures and their validation or contradiction by empirical observation and testing. By these standards, the basic Todaro and Harris-Todaro models as modified and refined by successive writers over the past fifteen years seem to have withstood the test of time reasonably well (see, for example, Robert Lucas, 1985, and

Fields, 1986). The Cole-Sanders critique of that model and the alternative framework which they offer is based on such restrictive and unrealistic assumptions that it does little to further our understanding of the migration process in developing nations.

REFERENCES

Bhagwati, Jagdish and Srinivasan, T. N., "On Reanalyzing the Harris-Todaro Model: Policy Rankings in the Case of Sector-Specific Sticky Wages," *American Economic Review*, June 1974, *64*, 502–08.

Cole, William E. and Sanders, Richard D., "Internal Migration and Urban Employment in the Third World," *American Economic Review*, June 1985, *75*, 481–94.

Corden, W. Max and Findlay, Ronald, "Urban Unemployment, Intersectoral Capital Mobility, and Development Policy," *Economica*, February 1975, *42*, 59–78.

Fields, Gary S., "Rural-Urban Migration, Urban Unemployment and Under-Employment, and Job-Search Activity in LDCs," *Journal of Development Economics*, June 1975, *2*, 165–87.

______, "Public Policy and the Labor Market in Developing Countries," in D. Newbery and N. Stern, eds., *Modern Tax Theory for Developing Countries*, World Bank, forthcoming, 1986.

Harris, John and Todaro, Michael, "Migration, Unemployment and Development: A Two-Sector Analysis," *American Economic Review*, March 1970, *60*, 126–42.

Johnson, George, "The Structure of Rural-Urban Migration Models," *East African Economic Review*, June 1971, *3*, 21–28.

Khan, M. Ali, "The Harris-Todaro Hypothesis and the Heckscher-Ohlin-Samuelson Trade Model: A Synthesis," *Journal of International Economics*, November 1980, *10*, 527–47.

Lipton, Michael, *Why Poor People Stay Poor: Urban Bias in World Development*, London: Temple Smith, 1977.

Lucas, Robert E. B., "Migration Amongst the Batswana," *Economic Journal*, June 1985, *95*, 358–82.

Mazumdar, Dipak, (1976a) "The Rural-Urban Wage Gap, Migration and the Shadow Wage," *Oxford Economic Papers*, November 1976, *28*, 406–25.

______, (1976b) "The Urban Informal Sector," *World Development*, August 1976, *4*, 655–79.

McCool, Thomas, "Wage Subsidies and Distortionary Taxes in a Mobile Capital Harris-Todaro Model," *Economica*, February 1982, *49*, 69–79.

Neary, Peter J., "On the Harris-Todaro Model with Intersectoral Capital Mobility," *Economica*, August 1981, *48*, 219–34.

Sabot, Richard H., "A Comparative Analysis of Rural-Urban Labor Market Interactions," mimeo., World Bank, 1975.

______, "The Meaning and Measurement of Urban Surplus Labour," *Oxford Economic Papers*, November 1977, *29*, 389–411.

______, *Migration and the Labor Market in Developing Countries*, Boulder: Westview Press, 1982.

Steel, W. F. and Takagi, Y. "The Intermediate Sector, Unemployment and the Employment-Output Conflict," mimeo., Department of Economics, Vanderbilt University 1976.

Stiglitz, Joseph, "Alternative Theories of Wage Determination and Unemployment in LDC's: The Labor Turnover Model," *Quarterly Journal of Economics*, May 1974, *88*, 194–227.

______, "The Efficiency Wage Hypothesis, Surplus Labour, and the Distribution of Income in L.D.C.s," *Oxford Economic Papers*, July 1976, *28*, 185–207.

Todaro, Michael P., "A Model of Labor Migration and Urban Unemployment in Less Developed Countries," *American Economic Review*, March 1969, *59*, 138–48.

______, "Reply to Comment on Labour Migration in Less Developed Countries," *Manpower and Employment Research in Africa*, No. 2, 1972, 49–51.

______, *Internal Migration in Developing Countries*, Geneva: ILO, 1976.

United Nations Economic and Social Council, Population Commission, Twentieth Session, *Concise Report on Monitoring of Population Policies*, E/CN.9/338, New York, 1978.

Internal Migration and Urban Employment: Reply

By WILLIAM E. COLE AND RICHARD D. SANDERS*

Despite the diversity and intensity of Michael Todaro's comment (1986), the essence of controversy reduces to differences in the respective views of the urban subsistence sector. Each view, in turn, depends upon a distinctive viewpoint. Observing from the viewpoint of the bureaucrat and other elite to whom he defers in his comment, Todaro sees the large cityward flow of Third World humanity as a menace, destroying urban amenities.

Todaro's focus being the modern sector, his model only incorporates an urban subsistence sector indirectly, and then only as a way station on the route to modern sector employment. His assumption is that every potential migrant has modern sector employment as an explicit goal.[1] Conceptually, he leaves no room for the uneducated rural persons whose aspirations are keyed to the modest employments of the urban subsistence sector. We, on the other hand, look from below, from the subsistence sector. From there, we see both rural and urban welfare gains from observed spatial movement of labor. We grant that unpleasant externalities may be involved, especially from the point of view of a discomfitted elite. However, those possible externalities should be studied in their own right and have no place in a debate between competing explanations of migration.[2]

The present debate, therefore, comes down to a straightforward empirical question. If, in fact, there are persons who move to the city with the intent of taking up permanent employment in the urban subsistence sector, then their decisions cannot be explained by Todaro. If they are few in number, the theorist may assume them away. If their numbers are significant, however, migration flows must be viewed as dual in nature and an alternative explanation for the subsistence portion is required.

Todaro's crucial point is that "the Cole-Sanders' theory requires them to make an artificial separation" of migration flows into those who go to the modern sector and those who go to and remain in the subsistence sector. This argument is effectively countered, we believe, by our crucial point that Todaro's view of an undifferentiated flow carries with it the inappropriate assumption that all migrants deem themselves able to enter into modern sector employment.[3]

*Department of Economics and Department of Statistics, respectively, University of Tennessee, Knoxville, TN 37996. This paper has benefited from comments by Walter C. Neale. We, however, retain responsibility for any errors.

[1] Todaro tells us there that there are three categories of migrants: There are those who go directly to the modern sector; those who get temporary jobs in the informal sector before moving on to the modern sector; and those who hope for modern employment, but who remain in fact in the informal sector. The category "informal sector" as used by Todaro would be a major component of what we have designated as the urban subsistence sector.

[2] Nevertheless, we yield to the temptation to point out that the restriction of automobile flows might protect urban amenities even more than would the restriction of human flows.

[3] Todaro, at one point, appears to think that the crux of our argument rests on the empirical relevancy of the simulation exercise reported in our Tables 1 and 2 (1985, p. 484). While we do not view the exercise as a crucial juncture, we stick with our numbers. We cite sources for our data that we believe to be as reputable as those he cites. As for Todaro's proposal of a discount factor of zero or possibly negative, we must demur. In the context of our discussion, a negative rate would be justified only if wages were rising as fast or faster than inflation, while interest rates lagged behind the inflation rate. For many *LDC*s, there is evidence that in the face of inflation, the purchasing power of blue-collar income declines. Todaro's negative discount rate would imply that the wage for a given job should provide for greater utility in the future than presently. In any event, the crucial question is not whether it takes two years or

Neither Todaro nor any of the empirical works cited by him has attempted to analyze data on this question.[4] We have. Where they have relied on aggregated data, usually of a census nature, we have turned to noted anthropologists and sociologists to obtain a view at the level of the household and the individual. It is to their works that we refer when we speak of data from many sources. And it is these works that Todaro disparages as "outdated, secondary Mexican statistics."[5]

One feature discussed by a number of social scientists, one of whom in the recent past was Todaro, is credentialism.[6] This phenomenon tells us not only that there are educational requirements for employment in the modern sector, but also that those requirements are often higher than would appear justified by the nature of the work. Information about these barriers to modern employment is widespread and potential migrants are likely to learn of them through networks that include relatives and acquaintances.[7] The evidence seems clear. A sizeable number of persons without significant relevant education, formal or informal, move to the cities. Far from being artificial, our conceptual separation of migration into two flows largely comports with reality.

We do not assume, as Todaro claims, that all persons who enter urban subsistence employment remain there permanently. Some migrants, those with sufficient education, accept temporary employment in the subsistence sector while awaiting selection into the modern sector. This aspect can easily be incorporated into our approach. We do admit to making simplifying assumptions, although not all of those alleged by Todaro, and none that are not common in highly aggregated models. We also concur that our explanation of migration between rural and urban subsistence sectors is simple and straight forward. In fact, it is straight from the textbook. Why then did we take such a round about way of getting to it? First, it was necessary to blow away some of the fog created by the Todaro model so that the reader could see that there is an important urban sector in which wage rather than employment equilibrates the labor market. Second, we think that our work goes beyond an explanation of the migration decision to explore the interrelationships between the modern and the subsistence sectors and to explore the determinants of demand for urban subsistence labor.[8]

twenty years for an educated migrant to get a modern job. The question is what happens to the migrant without the credentials required by the modern sector.

[4] Todaro chides us for failing to consider a number of works whose authors have sought to validate the Todaro model empirically. While not citing them, it is they to whom we refer when we say that his model has generally been found consistent with the data (1985, p. 492). We also add that we are counted among those who have fashioned econometric models to test the Todaro theme (1983). When Todaro cites several works as having developed more satisfactory models than ours for incorporating an urban subsistence or urban informal sector, we must again demur. Just as does Todaro, they consider the urban subsistence sector to be no more than a way station. Furthermore, they do not pursue intersectoral relationships and they especially overlook the determinants of demand for subsistence labor. Our own previous work (1972) is the earliest of which we are aware that treats both rural and urban subsistence sectors.

[5] A glance of titles in our original bibliography will tell the reader that our data covers more countries than Mexico. We now realize that we should have said that this paper "utilizes data" instead of "presents data." Possibly, then, Todaro would not have expected copious tables.

[6] In the latest version of his textbook, Todaro (1985) gives an extensive discussion of the problem of credentialism. Such awareness, however, is isolated from his discussion of internal migration.

[7] We find it incredible that Todaro can see all migrants as knowledgeable about urban wage rates, employment probabilities, and even able to arrive at a proper discount rate, while at the same time viewing them as ignorant of the existence of substantial barriers to modern sector employment.

[8] With reference to a sticky urban wage, Todaro wraps himself in the cloak of institutional realism and chides us for apparently overlooking the "facts of life" in *LDC*s. We acknowledge the fact of sticky wages in the modern sector but reject it as inappropriate for the subsistence sector. In their recent work, Allen Kelley and Jeffrey Williamson (1984, p. 19) agree with us that the wage equilibrates the subsistence labor market. Interestingly, they also independently arrived at the designation "*US*" for what we call the urban subsistence

Auditors' Report

February 26, 1986

Executive Committee
The American Economic Association

We have examined the balance sheets of The American Economic Association as of December 31, 1985 and 1984, and the related statements of revenues and expenses, changes in general fund and restricted fund balances and changes in financial position for the years then ended. Our examinations were made in accordance with generally accepted auditing standards and, accordingly, included such tests of the accounting records and such other auditing procedures as we considered necessary in the circumstances.

In our opinion, the financial statements referred to above present fairly the financial position of The American Economic Association as of December 31, 1985 and 1984, its revenues and expenses and the changes in its financial position for the years then ended, in conformity with generally accepted accounting principles applied on a consistent basis.

Touche Ross and Co.
Certified Public Accountants
Nashville, Tennessee

THE AMERICAN ECONOMIC ASSOCIATION BALANCE SHEETS, DECEMBER 31, 1985 AND 1984

	1985	1984
Assets		
CASH	$ 632,751	$ 915,479
INVESTMENTS, at market (Notes A and B)	4,327,785	3,165,188
ACCOUNTS RECEIVABLE, less allowance for doubtful accounts of $4,338 (1985) and $1,854 (1984)	108,933	101,250
INVENTORY OF *Index of Economic Articles*, at cost	100,860	90,918
PREPAID EXPENSES	12,846	22,383
OFFICE FURNITURE AND EQUIPMENT, at cost, less accumulated depreciation of $35,177 (1985) and $25,688 (1984)	61,300	40,676
	$5,244,475	$4,335,894
Liabilities and Fund Balances		
ACCOUNTS PAYABLE AND ACCRUED LIABILITIES	$ 528,843	$ 331,828
DEFERRED REVENUE (Note A):		
Life membership dues	44,436	47,058
Other membership dues	540,287	517,489
Subscriptions	449,263	417,104
Job Openings for Economists	19,867	18,138
	1,053,853	999,789
ACCRUAL FOR DIRECTORY (Note A)	68,758	211,610
FUND BALANCES:		
General	3,217,603	2,830,533
Unrecognized change in market value of investments (Notes A and C)	276,365	(147,997)
Net Worth	3,493,968	2,682,536
Restricted	99,053	110,131
Total Fund Balances	3,593,021	2,792,667
	$5,244,475	$4,335,894

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF REVENUES AND EXPENSES
FOR THE YEARS ENDED DECEMBER 31, 1985 AND 1984

	1985	1984
REVENUES FROM DUES AND ACTIVITIES:		
Membership dues and subscriptions	$ 831,233	$ 781,598
Nonmember subscriptions	614,155	608,416
Job Openings for Economists subscriptions	30,081	28,240
Advertising	107,835	102,404
Sale of *Index of Economic Articles*	56,121	12,433
Sale of copies, republications, and handbooks	26,522	28,122
Sale of mailing list	46,346	38,989
Annual meeting	15,651	20,958
Sundry	64,191	52,847
Operating Revenues	**1,792,135**	**1,674,007**
PUBLICATION EXPENSES:		
American Economic Review	610,132	475,735
Journal of Economic Literature	779,722	716,394
Directory publication (Note A)	50,000	65,000
Job Openings for Economists	53,910	49,859
Index of Economic Articles	45,444	9,854
	1,539,208	1,316,842
OPERATING AND ADMINISTRATIVE EXPENSES:		
General and administrative:		
Salaries	168,336	164,041
Rent	14,650	13,772
Other (Exhibit I)	190,503	162,197
Committee	41,728	53,711
Annual meeting	5,265	4,228
Benefit from federal income taxes (Note A)	(3,350)	(3,000)
	417,132	394,949
Operating Expenses	**1,956,340**	**1,711,791**
Operating Deficit	(164,205)	(37,784)
INVESTMENT GAINS (Note B)	449,361	284,557
REVENUES IN EXCESS OF EXPENSES	**$ 285,156**	**$ 246,773**

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN GENERAL FUND BALANCE

	Total	Operations	Market Value Adjustments
Balance at January 1, 1984	**$2,468,490**	**$1,620,985**	**$ 847,505**
Add market value adjustments resulting from inflation (Note A)	115,270	–	115,270
Add revenues in excess of expenses	246,773	246,773	–
Balance at December 31, 1984	**2,830,533**	**1,867,758**	**962,775**
Add market value adjustments resulting from inflation (Note A)	101,914	–	101,914
Add revenues in excess of expenses	285,156	285,156	–
Balance at December 31, 1985	**$3,217,603**	**$2,152,914**	**$1,064,689**

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN RESTRICTED FUND BALANCE

	Balance at January 1	Receipts	Disbursements	Balance at December 31
YEAR ENDED DECEMBER 31, 1984:				
The Alfred P. Sloan Foundation and Federal Reserve System grants for increase of educational opportunities for minority students in economics	$ 26,760	$141,570	$ 98,690	$ 69,640
The Minority Scholarship Fund for minority students applying for graduate work in economics	5,000	–	–	5,000
The Rockefeller Foundation Grant for minority students applying for graduate work in economics	51,780	890	22,680	29,990
Sundry	4,764	3,070	2,333	5,501
	$ 88,304	**$145,530**	**$123,703**	**$110,131**
YEAR ENDED DECEMBER 31, 1985:				
The Alfred P. Sloan Foundation and Federal Reserve System grants for increase of educational opportunities for minority students in economics	$ 69,640	$129,000	$ 93,193	$105,447
The Minority Scholarship Fund for minority students applying for graduate work in economics	5,000	–	–	5,000
The Rockefeller Foundation Grant for minority students applying for graduate work in economics	29,990	1,150	45,227	(14,087)
Sundry	5,501	100	2,908	2,693
	$110,131	**$130,250**	**$141,328**	**$ 99,053**

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN FINANCIAL POSITION FOR THE YEARS ENDED DECEMBER 31, 1985 AND 1984

	1985	1984
Cash, beginning of year	$ 915,479	$ 705,732
SOURCES OF CASH:		
Revenues in excess of expenses	285,156	246,773
Noncash charges:		
Depreciation	8,564	3,780
Directory publication (Note A)	50,000	65,000
Market value adjustments (Note A)	(187,415)	(20,262)
Cash provided by operations	156,305	295,291
INCREASE (DECREASE) IN CASH DUE TO CHANGES IN:		
Investments	(1,162,597)	62,299
Accounts receivable	(7,683)	73,537
Inventory of *Index of Economic Articles*	(9,942)	(42,926)
Prepaid expenses	9,537	(1,096)
Office furniture and equipment	(29,188)	(822)
Accounts payable and accrued liabilities	197,015	(10,811)
Deferred revenue	54,064	55,188
Accrual for directory	(192,852)	(531)
Restricted funds	(11,078)	21,827
General fund, market value adjustments	101,914	115,270
Unrecognized change in market value of investments	611,777	(357,479)
Cash, end of year	$ 632,751	$ 915,479

See notes to financial statements.

Notes to Financial Statements

A. Summary of Significant Accounting Policies

Investments are accounted for on a market value basis. According to the method the Association uses to value investments, the change in market value of corporate stocks, government obligations, bonds and commercial paper during the year, after adjusting for an inflation factor (3.3% in 1985 and 3.7% in 1984), is recognized in income over a three-year period for corporate stocks and reflected in current income for government obligations, bonds and commercial paper.

The Accrual for directory results because every three to five years the Association publishes a directory which lists, among other things, the names and addresses of its membership. This directory was most recently published in 1985 and distributed at no cost to the membership. In order to properly match the publishing cost of this directory with revenue from membership dues, the Association provided $50,000 in 1985 and $65,000 in 1984 for estimated publishing costs which will reduce actual directory expenses in the year of publication.

Deferred revenue represents income from membership dues and subscriptions to the various periodicals of the Association which are deferred when received. These amounts are then recognized as income following the distribution of the specified publications to the members and subscribers of the Association. Income from life membership dues is recognized over the estimated average life of these members.

The American Economic Association files its federal income tax return as an educational organization, substantially exempt from income tax under Section 501(c) (3) of the Internal Revenue Code. As required by Section 511(a) of this Code, the Association provides for federal income taxes on certain revenues which are not substantially related to its tax exempt purpose. This "unrelated business income" includes income from advertising and the sale of mailing lists. The Association has been determined to be an organization which is not a private foundation.

B. Investments and Investment Income

Investments consist of:

	December 31, 1985		December 31, 1984	
	Cost	Market	Cost	Market
Government obligations, bonds and commercial paper	$1,056,248	$1,176,371	$ 923,345	$ 983,784
Corporate stocks and mutual funds	2,146,227	3,151,414	1,696,178	2,181,404
	$3,202,475	$4,327,785	$2,619,523	$3,165,188

Investment gains recognized consist of:	Year Ended December 31	
	1985	1984
Government obligations, bonds, and commercial paper:		
Interest	$168,714	$176,304
Increase (decrease) in market value recognized	35,310	(28,084)
	204,024	148,220
Corporate stocks and mutual funds:		
Cash dividends	93,233	87,992
Increase in market value recognized (Note C)	152,104	48,345
	245,337	136,337
Investment gains, net	$449,361	$284,557

C. Unrecognized Change in Market Value of Investments

As described more fully in Note A, the Association recognizes in income over a three-year period changes in the market value of its corporate stocks. The following summarizes the years in which market value changes in stocks occurred that affected 1985 and 1984 revenues, and the amount of these market value increases (decreases) that will be recognized in income in future periods.

Year of Market Value Change	Recognized in Income in		To be Recognized in		Unrecognized Change	
	1985	1984	1986	1987	1985	1984
1982	$ –	$ 88,397	$ –	$ –	$ –	$ –
1983	69,747	69,747	–	–	–	71,600
1984	(109,799)	(109,799)	(109,798)	–	(109,798)	(219,597)
1985	192,156	–	192,156	194,007	386,163	–
	$152,104	$ 48,345	$ 82,358	$194,007	$276,365	($147,997)

The Association's revenues in excess of expenses would have been $709,520 in 1985 and ($130,968) in 1984 if changes in the market value of corporate stocks, after adjustment for inflation, had been recognized only, but entirely, in the year in which they occurred.

The Association's investment gains would have been $148,406 in 1985 and $121,978 in 1984 if 4% of the market value of the portfolio had been considered investment income (without regard to dividends, interest, capital gains and inflation). Under this method, the Association's revenues in excess of expenses would have been ($15,799) in 1985 and $84,195 in 1984.

D. Retirement Annuity Plan

Employees of the Association are eligible for participation in a contributory retirement annuity plan. Payments by the Association and participating employees are based on the employee's compensation. Benefit payments are based on the amounts accumulated from such contributions. The total pension expense was approximately $28,000 and $27,000 for 1985 and 1984, respectively.

E. Ratio of Net Worth to Expenses

The ratio of net worth at December 31, 1985 to 1986 budgeted expenses is 1.75 and the ratio of net worth at December 31, 1984 to actual 1985 expenses is 1.37.

EXHIBIT 1—THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF OTHER GENERAL AND ADMINISTRATIVE EXPENSES FOR THE YEARS ENDED DECEMBER 31, 1985 AND 1984

	1985	1984
Dues and subscriptions	$ 49,600	$ 49,600
Mailing list file maintenance	25,497	23,617
Postage	18,770	16,478
Accounting and legal	16,995	12,140
Periodic mailing expenses	15,054	14,189
Office supplies	14,823	11,988
Investment counsel and custodian fees	12,745	11,960
Currency exchange charges	9,516	1,548
Depreciation (straight-line method)	8,564	3,780
Telephone	5,489	4,626
Uncollectible receivables	4,772	1,002
Insurance and miscellaneous	4,060	4,088
President and president-elect expenses	2,941	5,648
Travel and entertainment	1,677	1,533
	$190,503	**$162,197**

ERRATA

Please note the following corrections and additions to the *AER Survey of Members*, December 1985, Number 6. (This page may be removed and inserted in the 1985 *Survey*.)

I. Corrections

P. 71– **ABELSON, MILTON,** ***Concurrent/Past Positions:*** Part-time consultant, John F. Kennedy Sch. of Govt., Harvard U., 1981–85; retired.

P. 205– **FRENKEL, JACOB A.,**
Phone: Office (312) 962–8253.

P. 214– **GENDREAU, BRIAN,** ***Degrees:*** B.A., Northwestern U., 1973; M.A., Sch. of Advan. Int'l. Studies, Johns Hopkins U., 1976.

P. 379– **NERLOVE, MARC L.,**
Phone: Office (215) 898–8484;
Home (215) 748–1924.

P. 406– **PHILPOT, GORDON,** ***Prin. Cur. Position:*** Prof. of Econs., Whitman Coll., 1983. ***Concurrent/Past Positions:*** Asst. & Assoc. Prof., Whitman Coll., 1969–82.

P. 432– **ROSEN, SHERWIN,**
Phone: Office (312) 962–8166.

II. Additions

HALDI, JOHN, 680 Fifth Ave., #1100, New York, NY 10019. ***Phone:*** Office (212) 664–8877; Home (212) 935–0050. ***Degrees:*** B.A., Emory U., 1952; M.A., Stanford U., 1956; Ph.D., Stanford U.,1958. ***Prin. Cur. Position:*** Pres., Haldi Associates, Inc., 1967–. ***Concurrent/Past Positions:*** Dir., PPB Staff, U.S. Bur. of Budget, 1965–67; Asst. Prof., Grad. Sch. of Bus., Stanford U., 1958–65. ***Research:*** Decision-making in business firms; strategic planning process.

KOOT, RONALD S., Pennsylvania State Univ., 310 Bus. Admin. Bldg., University Park, PA 16802. ***Phone:*** Office (814) 865–6861; Home (814) 238–1326. ***Fields:*** 210, 300. ***Birth Yr:*** 1937. ***Degrees:*** B.S., Penn. State U. at Univ. Park, 1962; M.A., U. of Oregon, 1967; Ph.D., U. of Oregon, 1967. ***Prin. Cur. Position:*** Prof. of Mngmt. Sci., Penn. State U. at Univ. Park, 1973–. ***Concurrent/Past Positions:*** Asst. & Assoc. Prof., Penn. State U. at Univ. Park, 1966–73. ***Research:*** Econometric analysis of macroeconomic stabilization policies.

NICHOLAS, CONSTANTINE J., P.O. Box 1752, Prince George Plaza Station, Hyattsville, MD 20902. ***Phone:*** Office (202) 475–3642; Home (301) 681–7179. ***Fields:*** 710, 530. ***Birth Yr:*** 1919. ***Degrees:*** B.S., U. of Pitt., 1941; M.A., U. of Pitt., 1949. ***Prin. Cur. Position:*** Econ., U.S. Dept. of Agric., Wash., D.C., 1965–. ***Concurrent/Past Positions:*** Traffic Mgr., Military Traffic Mgmt. Agcy., 1958–65. ***Research:*** New transport technologies in transporting agricultural products.

The Association does not keep corrections and additions on file for future editions of the *Survey*. Please do not send questionnaires to the Association office or to the data service company. Members in good standing at the time the next *Survey* is published will be asked to correct their listing or submit a new questionnaire.

NOTES

1987 *Nominating Committee of AEA*

In accordance with Section IV, paragraph 2, of the bylaws of the American Economic Association as amended in 1972, President-Elect Gary S. Becker has appointed a Nominating Committee for 1987 consisting of Charles L. Schultz, Chair; Marjorie Honig, Steve McGee, Frederic L. Pryor, Michael Rothschild, Richard Schmalensee, and Frank P. Stafford.

Attention of members is called to the part of the bylaw reading, "In addition to appointees chosen by the President-Elect, the Committee shall include any other member of the Association nominated by petition including signatures and addresses of not less than 2 percent of the members of the Association delivered to the Secretary before December 1. No member of the Association may validly petition for more than one nominee for the Committee. The names of the Committee shall be announced to the membership immediately following its appointment and the membership invited to suggest nominees for the various officers to the Committee."

Nominations for AEA Officers: 1987

The Electoral College on March 22 chose Robert Eisner as nominee for President-Elect of the American Economic Association in the balloting to be held in the autumn of 1986. Other nominees (chosen by the 1986 Nominating Committee) are: Vice President (two to be elected), Richard N. Cooper, Ann F. Friedlaender, Robert E. Lucas, Jr., and Burton A. Weisbrod; for members of the Executive Committee (two to be elected), Bernard E. Anderson, Robert J. Barro, Rudolph G. Penner, and Judith Thornton.

Under a change in the bylaws as described in the *American Economic Review Proceedings*, May 1971, page 472, additional candidates may be nominated by petition, delivered to the Secretary by August 1, including signatures and addresses of not less than 6 percent of the membership of the Association for the office of President-Elect, and not less than 4 percent for each of the other offices. For the purpose of circulating petitions, address labels will be made available by the Secretary at cost.

The Institute of International Education has awarded a grant to the American Economic Association under the Short-Term Enrichment Program (STEP) of the U.S. Information Agency to assist foreign graduate students studying in the United States to attend the annual meeting of the Association. Recipients must be full-time graduate students at U.S. institutions of higher learning and must not be receiving *any* U.S. government funds for academic or travel expenses. The 1986 meetings will be held in New Orleans, Louisiana, December 28–30. The maximum award is $300. To receive an application form, write STEP Grants, American Economic Association, 1313 21st Avenues South, Suite 809, Nashville, TN 37212–2786. To be considered for a travel grant, completed forms must be received by November 3, 1986.

The Social Science Research Council announces MacArthur Foundation Fellowships in Peace and International Security: 1986 Postdoctoral and Dissertation Training and Research Fellowships for scholars and students of any nationality or from any country. An award of $30,000 per year for postdoctoral fellows and $15,000 for dissertation fellows is provided to cover living expenses, travel, and research costs. The deadline is August 1, 1986. For full information, contact SSRC Program in International Peace and Security Studies, 605 Third Avenue, New York, NY 10158 (telephone 212+661–0280).

The Indo–U.S. Subcommission on Education and Culture offers twelve long-term (6–9 months) and nine short-term (2–3 months) awards for 1987–88 research in India. Applicants must be U.S. citizens at the postdoctoral or equivalent professional level. Fellowships include $1,500 per month ($350 payable in dollars; balance in rupees); an allowance for books, study/travel in India and international travel for grantee. There are allowances for long-term fellows' dependents. Application deadline in June 15, 1986. For full information and application forms contact CIES, Att: Indo-American Fellowship Program, Eleven Dupont Circle, NW, Suite 300, Washington, D.C. 20036–1257 (telephone 202+939–5469).

The Council for International Exchange of Scholars (CIES) announces the 1987–88 Fulbright Scholar Awards competition. Benefits include round-trip travel for the grantee and, for full academic year awards, one dependent; living costs allowance; tuition as well as book and baggage allowances. Grants are for 3–12 months in over 100 countries. Applicants must be U.S. citizens, hold the Ph.D. or comparable professional qualifications, and have university or college teaching experience. Deadlines are June 15, 1986 for Australasia, India, Latin America, and the Caribbean; September 15, 1986 for Africa, Asia, Europe, and the Middle East; November 1, 1986 for institutional proposals for the Scholar-in-Residence Program; January 1, 1987 for Administrators' Awards in Germany, Japan, and the United Kingdom; Seminar in German Civilization; and the NATO Research Fellowships; and February 1, 1987 for Spain Research Fellowships, and France and Germany Travel-Only Awards. For more information and appli-

cations, contact Council for International Exchange of Scholars, Eleven Dupont Circle, NW, Washington, D.C. 20036–1257 (telephone 202+939–5401).

Announcing the formation of the Korea-America Economic Association (KAEA): the objectives are to facilitate cooperation and participation in research efforts among economic professionals in all sectors who share interests in economic relations between North America and Korea. The KAEA will hold an international economic conference in August 1986 in Seoul, jointly with the Korean Economic Association. For full information on membership and KAEA activities, contact Taeho Bark, Department of Economics, Georgetown University, Washington, D.C. 20057 (telephone 202+625–4671/4121).

The Social Science Research Council announces an Advance Research Fellowship Program on the Processes of U.S. Foreign Policymaking, with support from the Ford Foundation. The fellowships will assist postdoctoral and senior scholars to undertake research on the processes of U.S. foreign policymaking and special emphasis will be given to proposals that seek to extend research in this area beyond the conventional focus on the foreign policy and national security agencies of the U.S. federal executive. Research is especially encouraged that compares the making of contemporary U.S. foreign policy processes to policymaking across historical periods, issues, or countries, as well as research that makes use of theories and insights from diverse social science disciplines. Awards will be for up to two years of full-time support and may be accepted in addition to other grants, awards, and fellowships. There are no citizenship or residence requirements. The size of the stipend is expected to range between $25,000 and $30,000 per year. The deadline is October 1, 1986. For full information, contact Council Fellowships in Foreign Policy Studies, SSRC, 605 Third Avenue, New York, NY 10158 (telephone 212+661–0280).

Announcing the creation of the International J. A. Schumpeter Society: The foundation will be held in Augsburg, West Germany, September 2–5, 1986, combined with an international congress on "Evolutionary Economics: Modelling and Empirical Research." The Society's aim is the scientific study of the problems of development, primarily in advanced economies. Following the ideas of Schumpeter, it conceives of development as the combination of growth and structural change broadly defined. It will concentrate on the dynamics of structural change, its origins and effects in all aspects: the role of the dynamic entrepreneur, income distribution, and will also consider political and social problems of entrepreneurship and entrepreneurial history. For further information, write to Professor Dr. Horst Hanusch, Lehrstuhl für VWL V, Universität Augsburg, Memminger Straße 14, D-8900 Augsburg.

Call for Papers: The annual meeting for the Association of Environmental and Resource Economists (AERE) will be held jointly with the AEA in New Orleans, LA, December 28–30. Those interested in having papers considered for this session should send three copies of a one-page abstract to the AERE President, V. Kerry Smith, Department of Economics, Vanderbilt University, Campus Box 52B, Nashville, TN 37235. The deadline is June 15, 1986.

Call for Papers: The ninth annual Middlebury College Conference of Economic Issues, "The Spread of Economic Ideas," will be held in the late fall, 1986. To be considered, please send proposals to David C. Colander, Johnson Professor of Economics, Middlebury College, Middlebury, VT 05753.

Call for Papers: *Studies in Economic Analysis* is a biannual, student-edited journal soliciting research articles from both established economists and students. Submission fee is $4.00 for nonsubscribers. Submit manuscripts to, or request complete format and style requirements from, The Editors, *SEA*, Department of Economics, College of Business Administration, University of South Carolina, Columbia, SC 29208.

Call for Papers: *Population Research and Policy Review* seeks manuscripts concerned with empirical research and public policy on topics relevant to population dynamics and structure. The journal wishes to emphasize the connection between research and policy and to focus on social issues implicating population, such as sex and race discrimination, urban programs, housing, immigration, and energy consumption. Submit two copies of papers to the Editor: Larry D. Barnett, School of Law, Widener University, P.O. Box 7474, Wilmington, DE 19803–0474.

Call for Papers: The fourteenth annual meeting of the Midsouth Academy of Economics & Finance will be held in Mobile, Alabama, February 25–28. Those interested in participating should submit abstracts and include areas that you would be willing to chair or discuss to Jim McMinn, MAEF General Program Chair, Department of Economics & Finance, Box 8537, Austin Peay State University, Clarksville, TN 37044. The deadline is October 15, 1986.

Call for Papers: The journal *Population Research and Policy Review* is planning a special issue on morbidity and mortality around the world. The co-editors for the issue are Charles B. Nam and George C. Myers. Manuscripts concerned with research on morbidity and/or mortality, that have policy implications, are being solicited. Papers should not exceed 20 double-spaced typewritten pages. Four copies of the paper should be sent to either Charles B. Nam, Center for the Study of Population, Florida State University, Tallahassee, FL 32306, or George C. Myers, Center for Demographic Studies, Duke University, 2117 Campus Drive, Durham, NC 27706. The deadline for submissions is late summer of 1986. For further information, write or telephone Dr. Nam (904+644–1762) or Dr. Myers (919+684–6126).

Call for Papers: The 1987 meeting of the History of Economics Society will be held June 20–22, 1987, at the Harvard Business School campus in Boston. Proposals for papers should include author(s) name, address, telephone number, and institutional affiliation; title and brief description of the paper. To be a discussant, indicate areas of interest. The theme of the meeting is the international scope of the interests and membership of the Society, but proposals and suggestions for the program are welcomed regarding any topic of interest to historians of economic thought and methodology. Proposals and suggestions should be sent by October 15, 1986 to the president-elect, Donald A. Walker, Department of Economics, Indiana University of Pennsylvania, Indiana, PA 15705.

Deaths

Alfred F. Chalk, professor emeritus, Texas A&M University, September 29, 1985.

Eric J. Hanson, professor emeritus of economics, University of Alberta, December 31, 1985.

Promotions

David L. Cleeton: associate professor of economics, Oberlin College, January 1986.

David T. Coe: head, General Economics Division, Department of Economics and Statistics, OECD, December 1985.

Victor J. Cook, Jr.: professor of marketing, A. B. Freeman School of Business, Tulane University, January 1, 1986.

Edwin Dean: chief, Division of Productivity Research, Office of Productivity and Technology, Bureau of Labor Statistics, December 1985.

Hillard G. Huntington: executive director, Energy Modeling Forum, Stanford University, September 1, 1985.

I. Curtis Jernigan, Jr.: chief, Economic Regulatory Section, Antittrust Division, U.S. Department of Justice, October 1, 1985.

Jon M. Joyce: chief, Economic Litigation Section, Antitrust Division, U.S. Department of Justice, October 1, 1985.

Youn-Suk Kim: professor of economics, Kean College of New Jersey, September 1, 1985.

Russell W. Pittman: assistant chief, Economic Regulatory Section, Antitrust Division, U.S. Department of Justice, October 1, 1985.

Frederick Warren-Boulton: Deputy Assistant Attorney General, Antitrust Division, U.S. Department of Justice, October 1, 1985.

Administrative Appointments

Joseph T. Bombelles; chairman, department of economics, John Carroll University, March 1985.

Morley Gunderson: director, Centre for Industrial Relations, University of Toronto, July 1, 1985.

Frank J. Navratil: dean, School of Business, John Carroll University, March 1985.

Richard C. Porter: chair, department of economics, University of Michigan, January 1, 1986.

Ranbir Varma: chair, economics department, Long Island University, Brooklyn Center, September 1, 1985.

New Appointments

Aquiles A. Almansi, University of Chicago: assistant professor, department of economics, University of Michigan, September 1, 1985.

Robert B. Barsky, MIT: assistant professor, department of economics, University of Michigan, September 1, 1985.

Charles C. Brown, University of Maryland: professor, department of economics, University of Michigan, September 1, 1985.

Christopher Garbacz, University of Missouri-Rolla: visiting professor of economics, National Taiwan University, February 1986.

Mary E. Fitzpatrick, Northwestern University: economist, Economic Regulatory Section, Antitrust Division, U.S. Department of Justice, December 1985.

Richard P. Rozek, Federal Trade Commission: senior analyst, Pharmaceutical Manufacturers Association, July 1, 1985.

Stephen W. Salant, Rand Corporation: professor, department of economics, University of Michigan, January 1, 1986.

Gim-Seong Seow: assistant professor of accounting, University of Washington, January 6, 1986.

George S. Tavlas, U.S. Department of State: senior economist, International Monetary Fund, September 1985.

Velma M. Thompson: associate professor of political economy and director of studies in public policy and free enterprise, School of Business and Management, Chapman College, September 1, 1985.

Mark L. Wilson: instructor of economics, University of Charleston, August 1985.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, use the following style:

A. Please use the following categories (please—do not send public relation releases):

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment: her new title (if any), new institution and the date at which the change will occur.

C. Type each item on a separate 3×5 card.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15; *December*, July 15.

All items and information should be sent to the Editorial Office, *American Economic Review*, 209 Nassau Street, Princeton, NJ 08542–4607.

NOTICE TO ALL GRADUATE DEPARTMENTS

The December 1986 issue of the *Review* will carry the eighty-third list of doctoral dissertations in political economy in American universities and colleges. The list will give recipients and titles of doctoral degrees conferred during the academic year terminating June 1986. This announcement is an invitation to send us information for the preparation of the list.

By June 30, please send us this information on 3×5 cards, conforming to the style shown below, one card for each individual. Please indicate by a classification number in the right-hand corner the field in which the thesis should be classified. The classification system is that used by the *Journal of Economic Literature* and printed in every issue.
Please note: Cards received that do not indicate a classification number will be printed as "Not Classified."

All items and information should be sent to the Editorial Office, *American Economic Review*, 209 Nassau Street, Princeton, NJ 08542–4607.

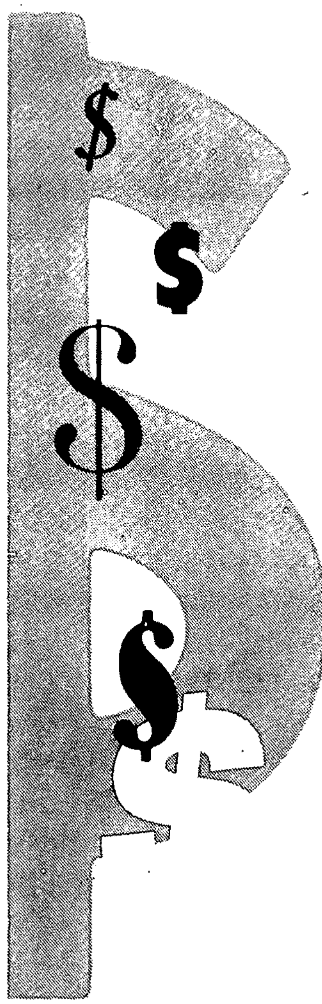
JEL Classification No. ________

Name: LAST NAME IN CAPS: First Name, Initial

Institution Granting Degree: ________________________

Degree Conferred (Ph.D. or D.B.A.) ____________ **Year** ____________

Dissertation Title: ________________________



SPREADSHEET STATISTICAL/FORECASTING PROGRAM ySTAT™ for Your IBM Personal Computer

WORKS LIKE LOTUS' 1-2-3, WITH FEATURES LIKE SAS

- **ySTAT is the only full-fledged spreadsheet statistical program** - a fast, powerful, easy-to-use stand-alone program.
- **Menu-driven format as in Lotus 1-2-3** - offers many functions that are not available even in 1-2-3, such as vector formulas, lagged variables, dummy variables, and row-column transposition.
- **You can read, enter, edit, copy, move, and sort the data** as well as group and partition the data for analysis right on the spreadsheet. A total of 40 function keys for one-keystroke operations.
- **Ten statistical function keys:** to tally sums and means; subtract variable means or seasonal means; standardize the data; aggregate seasonal data; generate seasonal indicators; perform moving averages, linear trend forecasts, and so on. An operation can be applied globally or locally.
- **Automatic recording feature** for programming macro keys, which with one keystroke, will replicate an entire sequence of analyses. You can save macros for future uses.
- **Verifiable data:** newly created and transformed data are immediately displayed.
- **ySTAT reads Lotus' files** and any free-format or fixed-format textfile generated by other PC programs or mainframe. Each record can have up to 168 variables. The data can be numeric, alphabetic, or a hybrid, including missing values.
- **No disk swapping:** ySTAT is compactly written - the entire program resides on one disk.
- **No output is too large:** a large analysis output that will not fit on the screen can be scrolled sideways, up, or down. With one keystroke, the output can be printed or saved as one piece.

STATISTICAL PROCEDURES

- **Summary statistics:** minimum, maximum, sum, mean, standard deviation, t-test, sum of squares, etc.
- **Frequency distribution:** frequency, cumulative frequency, percentage and cumulative percentage.
- **Correlation:** a matrix of correlation coefficients of all variables or selected variables.
- **Crosstab:** multi-way crosstabulation, including row percent, column percent, total percent and expected counts under the assumption of independence and the corresponding chi-square.
- **Analysis of variance:** one-way and two-way ANOVA including F-test for interaction effect.
- **Multiple regression (OLS):** up to 60 independent variables for 384K memory or more. Residuals are added to the spreadsheet for diagnostic analyses and plots.
- **Two-stage least squares (2LS):** for simultaneous equations and instrumental variable estimation.
- **Weighted least squares (WLS).**
- **Pooling of cross-section and time-series data:** regression analysis for deciding on whether or not to pool and to estimate the pooled regressions with different degrees of pooling.
- **Cochrane-Orcutt method:** an iterative method that includes the first observation in the estimation.
- **Two-sample Difference of means test:** by assuming equal variances as well as unequal variances.
- **Nonparametric tests:** the Kolmogorov-Smirnov test, the Wald-Wolfowitz runs test, the Mann-Whitney or Wilcoxon test, and the Wilcoxon matched- pairs signed-ranks test. Takes into consideration ties.
- **Spearman's correlation and Kendall's correlation.** Takes into consideration ties.
- **Plots:** standard plot, residual plot, time plot, paired plot, and X-Y scatter plot.

SPREADSHEET FORECASTING: Regression, exponential smoothing, and stochastic time-series methods to forecast and simulate nonseasonal time series and seasonal time series right on the spreadsheet.

- **Time-series models:** autoregressive model AR(p), moving average model MA(q) and **Box-Jenkins model** ARIMA(p,d,q), with autocorrelation and partial autocorrelation of residuals.
- **Nonseasonal forecasts:** linear trend, quadratic trend, polynomial model, sinusoidal model, exponential and double exponential smoothings.
- **Seasonal forecasts:** general exponential smoothing with seasonal indicators or trigonometric functions; Winter's additive and multiplicative procedures.

SYSTEM REQUIREMENTS: IBM Personal Computer (**PC, XT, AT**) or compatibles; 1 floppy disk or a hard disk; DOS 2.0 or 3.0 and 256K memory; with or without an 8087/80287 math co-processor chip.

PRICE: $395. Including one free update.

OUR GUARANTY: if you should find another statistical program easier to use than ySTAT, you may return ySTAT for a refund within 60 days from the date of purchase.

MING TELECOMPUTING INC.

Telecommunications and Statistics for Microcomputers

23 Oak Meadow Road, P.O. Box 101, Lincoln Center, MA 01773, U.S.A. (617) 259-0391

Please send: ( ) ySTAT Package for $395. (Massachusetts residents: please add 5% sales tax.)
( ) ySTAT Trial Disk and information/sample output for $5.
( ) ySTAT information/sample output.

System: ( ) IBM PC. ( ) IBM XT. ( ) IBM AT. ( ) other ____________
( ) with 8087 or 80287 math co-processor. Memory ________ K

Payment: ( ) a check is enclosed.
( ) Visa. ( ) MasterCard. Credit Card # ____________ Expiration Date: ___/___
( ) a university or governmental agency purchase order enclosed.

Name ____________ Signature (if charged) ____________

Address ____________

____________ Telephone: ( ) ____________

IBM is a registered trademark of International Business Machines Corp., 1-2-3 of Lotus, SAS of SAS Institute

Please mention The American Economic Review *When Writing to Advertisers*

The American Economic Review

ARTICLES

SEPTEMBER 1986

THE AMERICAN ECONOMIC ASSOCIATION

•Printed at Banta Company, Menasha, Wisconsin. The publication number is ISSN 0002-8282.

•*THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the *Survey*, and Supplements, is published by the American Economic Association and is sent to all members five times a year: March; May; June; September; December.

Regular member dues (nonrefundable) for 1986, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature* are as follows:

- \$37.50 if annual income is \$30,000 or less;
- \$45.00 if annual income is above \$30,000, but no more than \$40,000;
- \$52.50 if annual income is above \$40,000;
- \$18.75 annually for registered students (three years only).

Nonmember subscriptions will be accepted only for both journals: Institutions (libraries, firms, etc.), \$105 a year; individuals, \$70.00. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add \$12.00 to the annual rates above to cover extra postage.

•Correspondence relating to the *Survey*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

•Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

•Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786.

PAUL N. ROSENSTEIN-RODAN

DISTINGUISHED FELLOW

1985

Paul Rosenstein-Rodan graced the economics profession with deep analytical insights and practical wisdom, stimulating teaching and a humane concern for the welfare of the poorest people in the world. He was one of the founders and leaders of the field of development economics and his fundamental insights continue to shape its thinking. His contributions to the theories of value and money, which stressed the importance of complementarities and dynamic interactions, were related to his own intellectual origins in the Vienna of the late 1920's and London and Cambridge of the 1930's. In midcareer, in a major intellectual shift, he recognized and began his study of the problems of economic transformation and growth in poor countries. His original ideas were extended to become arguments for the need for sectoral and international balance, for intertemporal consistency, and an insistence on the importance of external economies. These, in turn, became the basis for his insistence on a program, rather than a project-by-project approach to development assistance. On the basis of his demonstrated concern with human welfare, his person integrity, and his evident intellectual grasp of development issues, he became the quintessential international economic confidant and advisor to presidents, prime ministers, and international agencies.

THE AMERICAN ECONOMIC REVIEW

- Submit manuscripts (4 copies), no longer than 50 pages, double spaced, to:
  Orley Ashenfelter, Managing Editor, *AER*; 209 Nassau Street, Princeton, NJ 08542–4607.
- Submission fee: $25 for members; $50 for nonmembers. Style guides will be provided upon request.
- No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

September 1986

VOLUME 76, NUMBER 4

Articles

Shorter Papers

Editorial Statement

It is the policy of the *American Economic Review* to publish papers only where the data used in the analysis are clearly and precisely documented, are readily available to any researcher for purposes of replication, and where details of the computations sufficient to permit replication are provided. The Managing Editor should be notified at the time of submission if the data used in a paper are proprietary, or if, for some other reason, the above requirements cannot be met.

Orley Ashenfelter
Managing Editor
Robert H. Haveman
John G. Riley
John T. Taylor
Co-Editors

Replication in Empirical Economics: The *Journal of Money, Credit and Banking Project*

By WILLIAM G. DEWALD, JERRY G. THURSBY, AND RICHARD G. ANDERSON*

This paper examines the role of replication in empirical economic research. It presents the findings of a two-year study that collected programs and data from authors and attempted to replicate their published results. Our research provides new and important information about the extent and causes of failures to replicate published results in economics. Our findings suggest that inadvertent errors in published empirical articles are a commonplace rather than a rare occurrence.

The confirmation of research findings through replication by other researchers is an essential part of scientific methodology. William Broad and Nicholas Wade in *Betrayers of Truth* (1983) present examples wherein the inability of other researchers to replicate published scientific findings revealed both inadvertent errors and outright fraud. Replications in the physical and social sciences are attempted infrequently, however. Thomas Kuhn (1970) emphasized that replication—however valuable in the search for knowledge—does not fit within the "puzzle-solving" paradigm which defines the reward structure in scientific research. Scientific and professional laurels are not awarded for replicating another scientist's findings. Further, a researcher undertaking a replication may be viewed as lacking imagination and creativity, or of being unable to allocate his time wisely among competing research projects. In addition, replications may be interpreted as reflecting a lack of trust in another scientist's integrity and ability, as a critique of the scientist's findings, or as a personal dispute between researchers. Finally, ambiguities and/or errors in the documentation of the original research may leave the researcher unable to distinguish between errors in the replication and in the original study. Months of effort may yield the replicator only inconclusive results regarding the validity of the original study, and thus no foundation for his future research in the area. These circumstances nurture a natural reluctance to undertake replication studies.

In July 1982, the *Journal of Money, Credit and Banking*, with financial support from the National Science Foundation, embarked upon the *JMCB Data Storage and Evaluation Project*. As part of the *Project*, the *JMCB* adopted an editorial policy of requesting from authors the programs and data used in their articles and making these programs and data available to other researchers on request. In a second part of the *Project*, we attempted replication of published results for a number of the submitted data sets. Our findings suggest that inadvertent errors in

*Dewald is Senior Economist, Bureau of Economic and Business Affairs, U.S. Department of State, Washington, D.C. 20520. Thursby and Anderson are Associate and Assistant Professors, respectively, Department of Economics, The Ohio State University, Columbus, OH 43210. We extend our thanks to the *Journal of Money, Credit and Banking* and to the present editors of the *JMCB* for their cooperation and decision to continue to request that authors submit data along with papers for review. We thank the National Science Foundation for financial support under contract SES-8112800. We extend special thanks to the authors who responded to our requests for programs and data, many of whom spent significant amounts of time and resources compiling materials for this project. We thank Hashem Dezbakhsh, Roger Lagunoff, and Joseph Hohman for research assistance, and Jeanette White, the *JMCB* secretary, for supervising the receipt and distribution of data sets. We thank Ed Kuh and MIT's Center for Computational Research in Economics and Management Science for loaning us the TROLL computer program. We also thank Benjamin Friedman, Lawrence Goldberg, Edward J. Kane, Thomas Mayer, Daniel Newlon, Edward J. Ray, Anthony Saunders, George Stigler, Stephen Stigler, Roger Waud, and Geoffrey Woglom for their comments and assistance. Responsibility for errors remains ours.

published empirical articles are a commonplace rather than a rare occurrence. While correction of the errors did not affect the authors' conclusions in most studies, the errors make independent replication impossible unless the replicator errs in precisely the same way.

In one crucial aspect, replications in empirical economics are simpler than in the experimental sciences: given the researcher's computer programs and data set, calculations may be repeated. The required programs and data are rarely available for replication, however. Many researchers employ proprietary statistical packages such as SAS, SPSS, TROLL, TSP, etc., which prohibit copying the program; replication is impossible for the individual researcher without subsidized (for example, major university) access to the same computer hardware and software. Many other researchers utilize programs which they or their research assistants have written in FORTRAN, Pascal, or other languages; interpretation and evaluation of these programs is difficult at best—and impossible at worst—without considerable skill, experience, and the cooperation of the original programmer. Dean Leimer and Selig Lesnoy (1982), for example, traced the false conclusions of Martin Feldstein (1974) to a computer programming error; we discovered similar errors in some of our replication studies. Finally, we note that some research projects employ computer programs of such enormous size and complexity as to all but guarantee that no other researcher will attempt replication of the study. The large-scale macroeconometric models such as the MPS model are members of this group. We discuss below our attempts at replication of a study based on the MPS model at Harvard University.

Similar problems arise with data. Some data are confidential, having significant proprietary value due to the difficulty and/or expense of their collection, while federal law makes other data available only for the internal use of employees of government agencies such as the Federal Reserve System. In the *JMCB Project*, many authors furnished their data even when the data had not been fully exploited in their own research. For other researchers, however, private interest prevailed and our request was either refused or ignored. We note that NSF Policy Number 754.2 *requires* that computer programs and data which have been produced with the assistance of NSF grants be made available to other researchers either by publication, duplication, or loan to the researcher. Investigators have the first right of publication, but the NSF rule requires that the programs and data be made available to others. It appears that this policy is seldom enforced and that investigators either are unaware of the policy or unafraid of the penalties for failure to comply with it.

I. The Role of Professional Journals

Professional journals disseminate authors' findings throughout the world. Our results suggest that journals take a more active role in assuring the quality of the results presented in empirical studies.[1] As editor of *Econometrica*, Ragnar Frisch argued for such a role in the first issue of that journal: "In statistical and other numerical work presented in *Econometrica* the original raw data will, as a rule, be published, unless their volume is excessive. This is important in order to stimulate criticism, control, and further studies" (1933, p. 3). Today, most economics journals except the *JMCB* do not have editorial policies which facilitate replication of published results by requesting programs and data sets from authors.

It is a matter of public record that errors exist in published empirical studies. Recent examples include Leimer and Lesnoy, cited above, and Frederick Siskind's 1977 correction of Finis Welch's 1974 minimum wage study. Our research suggests that there are many more unrecorded and undiscovered cases similar to these.

The frequency and magnitude of errors in empirical articles raise serious questions regarding the integrity of the refereeing process of professional journals. Referees are concerned primarily with methodology, the-

[1]An argument for the role of professional journals is presented by Edward Kane (1984).

oretical specification, statistical estimators, and importance of results; an author's programs, data, and calculations are typically assumed to be correct. While our findings suggest that this assumption often is unwarranted, we hesitate to suggest—due to the massive amount of time which would be required—that referees should be required to check an author's computer programs and data. Our findings suggest that the existence of a requirement that authors submit to the journal their programs and data along with each manuscript would significantly reduce the frequency and magnitude of errors. We found that the very process of authors compiling their programs and data for submission reveals to them ambiguities, errors, and oversights which otherwise would be undetected.

Our experience with authors who had not prepared programs and data for submission to the *JMCB prior to* submission of the article is indicative of the difficulties. Many could not locate the data for the article, while others had lost their programming. Even when the programs and data could be located, authors often had not kept a contemporaneous record of the progress of the research and could not reconstruct their results. In many cases—particularly in larger universities—graduate student research assistants had conducted essential parts of the research project, and after their departure it was impossible to reconstruct data sets from original sources.

In principle, the marketplace of economic research might be expected to provide a check against careless, undocumented empirical research. Since the editorial policies of professional journals had failed to address the problems of replication of published work, Edgar Feige proposed in 1975 that "...as a minimum standard, journal editors could explicitly publicize the necessity of full reporting of procedures and data..." (p. 1293). In response to Feige, the editors of the *Journal of Political Economy* wrote:

> We believe that the true remedy is resort to the powerful force of competition. We believe that journals should be prepared to accept alternative statistical tests of a hypothesis, in which either the confirmation or the contradiction of the author's statistical tests is reported. For this task to be reasonably economical, any author should be willing to provide his underlying data to other scholars (at cost). Indeed, this behavior is a requirement for responsible scholarship.
>
> [1975, p. 1295]

The editors subsequently added a new section to the journal devoted to verifications and contradictions of papers first published in the *JPE*. Invariably, this section contained papers employing either new data sets or alternative statistical techniques; little attention was paid to replication. Further, the *JPE* neither required nor facilitated making programs and data available for replication attempts.

The *JPE* experiment is a classic example of market failure. The benefits of reduced frequency and magnitude of errors in empirical articles share many of the characteristics of public goods: all who read the journal benefit from the knowledge that the research reported in its articles has been more carefully monitored by the researcher; the quantity of benefits available to any single reader is not reduced by others reading the journal; and it is difficult to induce the reader to reveal his or her true value (price) for better quality articles. A single researcher faces high costs in time and money from undertaking replication of a study and finds no ready marketplace which correctly prices the social and individual value of the good.

An editorial policy which requires the submission of programs and data to the journal has two significant advantages relative to a *laissez-faire* system wherein interested researchers must contact authors directly. First, it substantially reduces the cost borne by a researcher seeking to replicate original research. The economic self-interest of the author in satisfying the editors of the journal assures that the materials submitted to the journal are more complete and correct than what if anything might be furnished to an individual researcher. The journal provides a cost-effective clearinghouse for these materi-

als, relieving the author of the burden of furnishing programs and data to many researchers individually. Perhaps most importantly, potential replicators avoid giving the impression of challenging an author's results by obtaining programs and data from the journal rather than from the author. Thus, these editorial policies reduce barriers-to-entry in replication and thereby facilitate the market-driven outcome suggested by the editors of the *JPE*. Second, readers of the journal benefit by knowing that the likelihood of an error in a published study has been reduced by careful preparation of the programs and data for submission to the journal, and that the referees of the article have had access to the programs and data. In this respect, the editorial policy is a form of professional collective action which solves the public goods problem by combining the values of the good to individual readers.

In recognition of the importance of data collection and construction, the *Review of Public Use Data* and the *Journal of Econometrics* have entered into an arrangement whereby authors of applied articles in the *Journal* are given preference to publish supplementary data in the *Review*. Overall, however, professional journals in economics have not adopted editorial policies to facilitate replication, and there have been few attempted replications.

II. The Response of Authors to Requests for Data

The *JMCB Project* requested programs and data from the authors of all empirical articles published during or after 1980. These requests are usefully divided into two groups. The first group consists of authors of articles published prior to the start of the *JMCB Project* in July 1982. This control group had submitted and published articles in the *JMCB* without knowledge that a subsequent request would be made by the *JMCB* for their programs and data. The second group consists of authors whose articles, beginning July 1982, were either: (*i*) accepted for publication but not yet published, or (*ii*) under review by referees. We are grateful to the many authors who supplied programs and data, thereby prospectively subjecting their research to replication. For others, the nonresponse rate is itself an important finding of the *Project*. Table 1 summarizes the responses of authors to our requests.

In the first group, 42 of 62 authors responded to our request and 22 of the authors submitted programs or data. Approximately one-third of the authors (20) never replied to our repeated requests, and an additional one-third (20) replied that they could not furnish their programs or data. The motives of the 20 authors who did not reply to our requests are not known. Much more informative are the responses of the 20 who replied but did not furnish programs and data. Two authors wrote that their data were confidential and could not be released. Fourteen wrote that they had lost or discarded their data. The others wrote that their data were readily available from published sources but did not furnish the data, leaving collection of the data to us. We report below the results of an experiment wherein we attempted to replicate one study by collecting data from published sources.

Why were 18 of these 20 authors unable to provide programs and data, some as little as six months after publication of the article? We surmise that some authors simply may not keep programs and data after completion of a research project and, in any case, that authors devote most of their effort to the completion of a publishable manuscript and little to the tedious task of compiling, rechecking, and documenting programs and data. The authors who submitted data informed us that compiling the materials which we requested was often a lengthy and expensive task. A low-cost alternative is to ignore the request for data as just another questionaire, or to reply that the data are lost, destroyed, or available from published sources. Economists recognize that optimizing behavior allocates resources to their most productive uses. Our results are not unexpected when professional rewards generally arise from the publication of articles and rarely from documentation of the research or providing the underlying materials to other researchers.

We emphasize that authors in this first group were not aware during the course of

TABLE 1—RESPONSES TO REQUESTS FOR DATA FROM AUTHORS OF EMPIRICAL PAPERS[a]

	Published before Data Requested	Accepted before Data Requested	Under Review when Data Requested
Requests	62	27	65
Responses	42	26	49
Response Rate (Percent)	66	96	75
Mean Response Time (Days)	217	125	130
Not Submitted:			
Confidential Data	2	1[b]	0
Lost or Destroyed Data	14	2	1
Data Available, But Not Sent[c]	4	2	1
Nonrespondents	20	1	16
Total Not Submitted	40	6	18
Nonsubmission Rate (Percent)	66	22	28

[a] Includes all requests made through December 1984, and excludes authors whose papers were rejected.
[b] Two data sets were partially confidential.
[c] This category includes authors who (*i*) stated that their data were available from published sources, but did not send their data; and (*ii*) authors who claimed to have their data but were unwilling to sort through their papers to find the data.

their research that the *JMCB* would subsequently request their programs and data. Several authors wrote that they easily could have supplied programs and data at the time of initial submission of the manuscript for review. Nevertheless, it is surprising that they did not retain their programs and data for a year or two after publication, not only for their own research but also, where research was supported by NSF, to satisfy its policy requiring that such information be provided to other researchers.

The responses of the second group of authors are summarized in the second and third columns of Table 1. The proportion of authors who submitted programs or data is significantly larger than the 34 percent submission rate of authors in the first group: 72 and 78 percent for papers-under-review and papers-accepted-but-not-published, respectively. The reasons for not submitting programs and data were similar to those given by the first group of authors. Two authors of accepted-but-not-published papers reported that they had already lost or destroyed their data, and one of the 27 authors in this group never replied to our repeated requests for data.

The response rate of authors whose papers were under review by referees was only 75 percent. All authors were given a minimum of six months to respond, and nonresponse was followed by a second request. While the *JMCB* has not made submission of the data a requirement for either submission or publication of a paper, it is nonetheless noteworthy that one-fourth of the authors would not even reply to a letter from the journal reviewing their manuscript for publication. We can speculate that some authors did not compile or organize their programs and data prior to submission of their manuscript and would have responded if a favorable publication decision was reached, but perhaps not. One of the authors who replied said that he had already lost or destroyed the data before a decision had been reached regarding publication of the manuscript.

III. Characteristics of the Submitted Datasets

Data are useless to another researcher unless accurately recorded and properly documented. Our goal for each submitted data set was that it be complete enough to allow an attempt at replication. We examined the first 54 submitted data sets to determine how often that goal was met. Eight, or 15 percent, were judged to have met that goal, while 14, or 26 percent, were judged incomplete. The

TABLE 2—PROBLEMS IN SUBMITTED DATA SETS

	Published before Data Requested	Accepted before Data Requested	Under Review when Data Requested
No Problems	1	3	4
Problems Identified:			
Incomplete Submission	6	3	5
Sources Cited Incorrectly	0	4	4
Sources Cited Imprecisely	11	7	10
Data Transformations Described Incompletely	3	4	1
Data Element Not Clearly Defined	2	3	2
Other	0	3	1
Problems	22	24	23
Data Sets Examined	19	14	21

problems which we encountered are summarized in Table 2. The most frequent problem was a failure to identify the sources of the data precisely. This problem was somewhat more common among data sets submitted by authors in the first group who, in some cases, had completed the research project several years earlier. Similarly, incomplete data sets also were more common with this group. Further, only 34 percent of the authors in this group were willing or able to submit any data.

The identification of individual variables was a problem in many data sets. Submitted programs and data sets often were so inadequately documented that we could not identify the variables which had been used in calculating the published empirical results. The variable names in some data sets did not match those in the published articles, some data sets contained no variable names at all, and a few data sets omitted the original data from which the author had constructed new transformed variables. In addition, some authors had discarded data on variables which they reported as having insignificant coefficient estimates in their regressions. While we attempted to resolve any ambiguity by contacting authors, usually they had sent us all available information and the ambiguity could not be resolved.

We attempted to replicate several data sets from the authors' stated original published sources. Many authors cited only general sources such as *Survey of Current Business*, *Federal Reserve Bulletin*, or *International Financial Statistics*, but did not identify the specific issues, tables, and pages from which the data had been extracted. Since government economic data may be revised several times after their initial publication, we often found ourselves unable to reconstruct data sets from such vague documentation. We present in Section V an experiment which illustrates the wide range of possible results that may be obtained in replications based on published data.

Time and resources did not permit replication and reconstruction of all submitted data sets. Detailed examination of a sample

TABLE 3—CANARELLA AND GARSTON ORIGINAL AND CORRECTED RESULTS, LIKELIHOOD-RATIO TESTS OF RATIONAL EXPECTATIONS AND DEBT NEUTRALITY, REAL OUTPUT MODEL

Hypothesis Tested	Original Results	Revised Results
REH and EH Restrictions Applied to		
Money	6.246	7.014
Debt	2.302	.774
Money and Debt	8.286	8.632
SN and EH Restrictions Applied to		
Money	6.350	15.808[b]
Debt	2.794	4.274
Money and Debt	13.364[b]	20.930[b]
REH, SN and EH Restrictions Applied to		
Money	12.454	22.802[a]
Debt	4.272	4.594
Money and Debt	20.744[a]	28.888[b]

Note: *REH* = Rational expectations hypothesis; *EH* = Efficiency hypothesis; *SN* = Structural neutrality.
[a] = Significant at 5 percent level.
[b] = Significant at 1 percent level.

TABLE 4—MAYER AND NATHAN ORIGINAL AND CORRECTED RESULTS, CONTRACTED RATE ON NEW HOME MORTGAGES ORIGINATED BY MUTUAL SAVINGS BANKS REGRESSED ON YIELDS ON TEN-YEAR U.S. GOVERNMENT BONDS[a]

	Ten-Year Government Bond Rate				
Equation	Current	1-Year	R^2	*D-W*	*Rho*
Original Results					
7	.819		.987	2.3	−0.170
	(14.6)				(−.6)
8		.562	.939	2.2	.186
		(3.7)			(.7)
Corrected Results					
7	.887		.981	2.8	−.524
	(15.2)				(−1.6)
8		.020	.876	.9	.020
		(3.4)			(−.03)

[a]The *t*-values are shown in parentheses.

of the data sets revealed a number of data errors, some of which significantly changed the statistical results and conclusions of the studies. Three examples illustrate our findings. We discovered a number of transcription errors in the data set submitted by Giorgio Canarella and Neil Garston (1983) before the manuscript was sent to the printer. While their conclusions are unchanged, correction of the errors caused significant changes in the reported regression estimates and likelihood-ratio test statistics. A sample of the original and published results is shown in Table 3.

Our examination of the data set submitted by Edward Gramlich (1983) revealed a conflict between the sample periods cited in the manuscript and in the data set. While checking this problem, the author discovered that one of the forecasts had been coded incorrectly and was out-of-phase with the other forecast by six months. Correction of this problem substantially changed some conclusions regarding the relative accuracy of the forecasts. The corrected results appear in the published article.

Finally, a minor coding error in the ten-year government-bond-rate series used by Thomas Mayer and Harold Nathan (1983) was discovered after the *JMCB* had gone to press; an "Errata" was subsequently published (1984). While their conclusions were unchanged, the corrected equations are quite different (see Table 4).

IV. Summary of Replications from Submitted Datasets

We conducted replications of nine articles from the *JMCB* as part of the *JMCB Project*. Our choice of articles was limited to those by authors who submitted data sets to the *JMCB*, and we extend our thanks to these authors for exposing their work to our replication efforts. We believe the articles are representative of the *JMCB* in terms of content and econometric sophistication. The number of replications was limited, and we did not in every case attempt to replicate all the results of the article. Our goal was to obtain the same numerical results as had been obtained by the authors and not necessarily to determine whether those results would be confirmed by further scrutiny such as checking the submitted data against original sources. While it is possible that we have erred in these replications, we have made extraordinary efforts to reduce the likelihood of such errors, including contacting the authors to discuss their articles and our replication findings.[2] The replication results are summarized in Tables 5–8.

We replicated the results of two articles in their entirety: James Johannes and Robert

[2]We do not mean to suggest that all authors agree with our findings, only that we have endeavored to keep them informed of our replication efforts.

Rasche (1981) and Robert Engle (1983). Replication was aided greatly by the exceptionally clear, detailed programs and data sets submitted by the authors. Even in these cases we encountered minor problems due to the use of different computer hardware and software. The problems were resolved quickly by contacting the authors. Similar problems appear inherent in replications even when the authors furnish excellent descriptions of their programs and data.

We reproduced exactly almost all the results of V. Vance Roley (1983) and obtained qualitatively similar results for John Merrick (1983). Roley tested the impact of weekly money stock announcements on Treasury bill yields by estimating a three-equation model over each of three sample periods. We were able to replicate from his data set all of Roley's results except his estimates of the third equation (equation (3), p. 350) for the first sample period (see Table 5). Merrick estimated a single-equation model of the determinants of money growth (Table 1, p. 227). Despite our best efforts, we were unable to reproduce exactly his regression estimates. Our regression coefficients do display the same sign and statistical significance as those reported by Merrick except for the coefficient of R^S_{t-2}: our coefficient estimate is negative and insignificant (a value of -0.0001 with a t-statistic of 0.01), while Merrick reports a value of 0.029 with a t-statistic of 4.83 (see Table 6). The reasons for these differences are unknown.[3]

We discovered several computer programming errors in our replications, some minor, some serious. A typical example is Brian Maris (1981). The article cites 1952:III–1977:III as the estimation period (101 observations) while the initial period used for estimation in the computer programs is 1950:III (110 observations); Maris's computed Box-Pierce χ^2 statistics are calculated from 101 residuals, not 110. The FORTRAN program used for estimation erred by attempting to read a number from a memory location beyond the end of an array. This FORTRAN error forces the computer to interpret whatever information is stored in that memory location as a real number, and produces highly unpredictable results. Using

TABLE 5—ESTIMATES OF THE EQUATION NOT REPLICATED FROM ROLEY (1983)[a]

Coefficient	Roley	Our Estimate
β_0	0.0029(.0079)	0.0003(.0085)
β_{10}	−0.0094(.0178)	−0.0025(.0152)
β_{11}	0.0059(.0061)	−0.0030(.0049)
β_{20}	0.0463(.0303)	0.1113(.0475)
β_{21}	−0.0125(.0152)	−0.0476(.0257)
β_{30}	0.0472(.0275)	−0.0312(.0213)
β_{31}	0.0245(.0084)	0.0156(.0057)
β_{40}	0.0275(.0117)	0.0302(.0146)
β_{41}	−0.0073(.0030)	−0.0094(.0041)
β_{50}	−0.0098(.0545)	0.0160(.0374)
β_{51}	−0.0045(.0272)	−0.0067(.0154)
β_{60}	0.0296(.0165)	0.0253(.0192)
β_{61}	−0.0093(.0054)	−0.0085(.0068)
$\bar{R}^2$	0.19	0.19
SEE	0.036	0.035
$D\text{-}W$	1.94	2.05

[a]Standard errors are shown in parentheses; $D\text{-}W$ = Durbin-Watson statistic.

TABLE 6—ESTIMATES OF MONEY GROWTH RATE EQUATION FROM MERRICK (1983)[a]

Coefficient	Merrick	Our Estimate
a_0	0.007(1.40)	0.013(2.41)
a_1	0.606(7.97)	0.626(6.85)
a_2	−0.054(−0.58)	−0.039(−0.38)
a_3	0.116(1.27)	0.141(1.40)
a_4	−0.071(−0.78)	−0.047(−0.46)
a_5	0.097(1.24)	0.099(1.14)
a_6	0.075(1.07)	0.065(0.85)
a_7	0.008(2.31)	0.010(2.65)
a_8	−0.005(−1.09)	−0.002(−0.49)
a_9	0.011(1.61)	0.011(1.45)
a_{10}	−0.003(−0.76)	−0.004(−0.93)
a_{11}	−0.273(−3.14)	−0.380(−4.06)
a_{12}	0.306(3.64)	0.361(3.90)
a_{13}	0.029(4.83)	−0.0001(−0.01)
$D\text{-}W$	1.98	1.98

[a]The t-statistic values are shown in parentheses.

[3]One reason may be the differing abilities of numerical algorithms in econometrics computer programs to cope with collinear time-series data. Our results were obtained with version 3.4B of the TSP Econometrics package written in double-precision FORTRAN on an IBM 3081D and also with regression programs written by the authors of this paper in double-precision IBM VS FORTRAN using high-accuracy algorithms from the IMSL computer subroutine library. We believe that authors should investigate the numerical stability of their results and publish the findings.

TABLE 7—BATAVIA AND LASH (1982): PUBLISHED AND REPLICATED RESULTS, *GNP* REGRESSION

	Constant	Coefficient *M*	*H*	*L*
A. ***GLS Results***				
Batavia and Lash	−.8116	1.4623	.0722	.2970
	(3.98)	(46.62)	(1.75)	(2.59)
Replication:				
Iterative	.1561	1.2857	.0169	.1736
Method	(.331)	(15.068)	(.471)	(1.399)
Replication:				
Single	−.7533	1.4522	.0708	.312
Iteration	(−3.79)	(47.41)	(1.781)	(2.786)
B. ***2SLS***				
Batavia and Lash	.7496	1.4541	.0522	.3358
	(2.34)	(31.31)	(1.30)	(1.61)
Replication:				
Iterative	−.4510	1.3965	.0294	.2775
Method	(−1.408)	(24.66)	(.792)	(1.859)
Replication:				
Single	−.697	1.444	−.0542	.3429
Iteration	(−3.003)	(40.48)	(1.409)	(2.519)

Note: The t values are shown in parentheses. M = Money supply; H = Ratio of high employment government expenditures to high employment tax receipts; L = Ratio of bank loans to bank earning assets.

the time-series filters reported by Maris, we were able to replicate all the results in his article. After correcting the FORTRAN program, we found that the specific time-series filters accepted by Maris were rejected by the data. We then identified and estimated an alternative set of filters which yielded innovations that passed the Box-Pierce test; our corrected results support Maris's causality conclusions.

We could not replicate the results of two articles, in one case even with the active assistance of the author. In Bala Batavia and Nicholas Lash (1982), the authors state that generalized least squares was used for the estimation of both a single-equation and a simultaneous-equation model, but do not present the estimator. The authors were unable to furnish their computer programs and we were unable to replicate the authors' regression estimates from the data set which they furnished. We obtained estimates qualitatively similar to their estimates in magnitude and statistical significance by using a single-iteration Cochrane-Orcutt estimator with the value of ρ fixed at the value reported in the article. The values and significance levels of coefficient estimates obtained from both single-iteration and iterative Cochrane-Orcutt estimators differed greatly from those reported by the authors (see Table 7).

The second replication examined Geoffrey Woglom's 1981 study of the role of stock prices as a determinant of consumption in the MIT-Penn-SSRC (MPS) macroeconometric model. The research reported in the article had been completed several years prior to our request and the data set had been updated since publication of the article; the vintage data set which corresponded to the article itself could not be reconstructed. Woglom furnished his current data set which included revised observations on the variables and time periods used in the 1981 article. Both we and Woglom attempted replication of the article so as to determine whether the loss of the vintage data was important. While ours and Woglom's estimates based on the revised data are very similar, both differ from those contained in his article. Table 8 presents results for his equations 7 and 9 (Table 1, p. 218). Estimates based on the revised data set generally are more consistent than his published estimates with the hypothesis that the non-

TABLE 8—WOGLOM (1981): PUBLISHED, REVISED, AND REPLICATED ESTIMATES OF CONSUMPTION EQUATIONS

	YD	*U·YD*	*VKN*	*FVST*	*RVST*	*F*-Stat for Test of *e* > *c*
Equation 7 (1963:I–1977:III)						
Original Estimates	.694	−.073	84.5	−38.6	32.2	3.78
	(18.7)	(−.4)	(6.3)	(−2.1)	(1.4)	
Woglom's Replication	.78	−.099	49.75	−11.64	65.65	[a]
	(30.0)	(−1.01)	(5.29)	(−1.12)	(3.83)	
Our Estimates	.716	.177	61.59	−10.51	150.23	4.23
	(12.31)	(.819)	(3.42)	(−1.19)	(2.38)	
Equation 9 (1963:I–1972:IV)						
Original Estimates	.744	−.041	81.9	−75.5	6.6	10.98
	(39.8)	(−.3)	(8.8)	(−3.3)	(.4)	
Woglom's Replication	.777	.198	57.94	−51.65	26.16	[a]
	(27.75)	(.938)	(5.51)	(−1.51)	(1.39)	
Our Estimates	.742	.484	65.02	−50.19	292.34	6.20
	(35.35)	(4.17)	(6.42)	(−5.32)	(6.76)	

Note: The *t*-statistics are shown in parentheses. *YD* is disposable income, *VKN* is the value of nonstock, household net worth, *FVST* is the value of common stocks due to quantifiable factors, and *RVST* is the value of common stocks due to nonquantifiable factors. *U* is a measure of the unemployment rate. The coefficients on *YD*, *U·YD*, *FVST*, and *RVST* are the sums of lag coefficients on second-degree polynomial lags constrained to zero at the tail. The lags on *YD* and *U·YD* are 9 periods (including the present period) and the lags on *FVST* and *RVST* are 7 periods (including the present period).

[a] Not reported.

quantifiable value of stocks reflects expectations of future income. These results emphasize the importance of maintaining intact the vintage data sets used in published articles, especially when continuing research requires that active data sets be updated with revised observations.

Several authors wrote that all or part of their data had been lost but could readily be obtained from published sources. We chose to replicate the results of a typical article in this group (Lawrence Goldberg and Anthony Saunders, 1981). Goldberg and Saunders provided the banking data used in their article but their data on imports, investment, and *GNP* had been lost. They identified the general source for each variable (either the *Survey of Current Business* or *Federal Reserve Bulletin*), but provided no specific months, pages, or table numbers. They stated in materials submitted to the *Project* that they estimated the system of equations in their article by generalized least squares; in subsequent conversation, they explained that the agencies equation was estimated by ordinary least squares and the other two equations by a single-iteration Cochrane-Orcutt estimator. During December 1982 we collected the most-recently published values for all variables and time periods in their model.[4] Our calculated numerical values for the regression coefficients and standard errors differed significantly from those published by Goldberg and Saunders, with some coefficients which had insignificant *t*-statistics in their article becoming significant in the replication and vice versa. We examine this article further in Section V below.

The final replication study which we discuss here concerns a large-scale macroeconometric model. In 1982, the *JMCB* published an article by Benjamin Friedman based on a version of the MIT-Penn-SSRC (MPS) macroeconometric model containing a large model of the market for U.S. government bonds. We requested the programs and data used for the article, appreciating that this could be a complex task. To our knowledge, no one previously had addressed the feasibility of a researcher furnishing the programs

[4] Goldberg and Saunders cited the *Federal Reserve Bulletin* as the source for some data series which also are published in the *Survey*. Since the *Survey* is the primary source for these series, we collected all data from the *Survey*.

and data from a large macroeconometric model to another researcher for replication of an article. In July 1983, the author sent us an 87-page manual describing the installation and usage of the MPS model on Harvard University's IBM VM/370 computer system (Friedman and David Johnson, 1983), and two computer tapes containing more than 2500 files of programs and data.

Our problems in using Friedman's programs and data were the familiar difficulties of moving programs and data across computer systems, complicated by the scale of the MPS model. An accurate appraisal of the time, knowledge, and resources required of another researcher seeking to replicate Friedman's results only could be obtained by installing his MPS model on a computer, in this case, Ohio State University's IBM 4341 VM/CMS system. The Harvard-MPS model system includes a supervisor program written in the EXEC language of IBM's CMS operating system; a set of FORTRAN programs to define the model's equations, load data, and solve the model in simulation; and the NBER TROLL econometrics package to store, retrieve, and analyze simulation output. The model's FORTRAN programs are not integrated into the TROLL system, and communications between them, TROLL, and the user are conducted by the CMS EXEC program handling numerous disk files. We found installation of the TROLL package on the IBM 4341 relatively straightforward, but not a task to be attempted without prior experience with IBM VM computers.

Our most difficult task was converting the CMS EXEC programs from IBM 308X series mainframe computers using IBM 3350-type disk drives to an IBM 4341 CMS computer using 3370-type disk drives. Our limited time, patience, and resources forced us to end this replication effort before we finished the necessary reprogramming. In a subsequent telephone conversation, Friedman estimated that the time and effort required to produce the manual and computer tapes approximated that required to produce an additional article for a professional journal (these materials were compiled specifically in response to our request), and admitted astonishment that anyone would attempt to convert the Harvard-MPS-TROLL system to another computer and use it for replication.

Replication attempts which use large-scale eonometric models are arguably the most complex of all such attempts due to the large computers, programs, and data bases required. A researcher without an IBM VM/CMS computer system would find replication impossible since the programs are fully IBM dependent. As our replication effort demonstrates, even installation on another IBM computer may require a substantial amount of technical expertise (or the research grants to hire it). Nevertheless, the desirability of scientific replication is—if anything—greater with respect to results based on complex computer technology than for smaller more-easily reproducible models.

V. The Importance of Vintage Data in Replication: A Simulation Experiment

A number of authors argued that their data "...could be readily obtained from published sources" and submitted no data or only partial data sets for the *JMCB Project*. Government agencies periodically revise published data such that the value of a variable for a particular time period (for example, gross domestic investment in 1978:II) differs in different issues (volume, number) of the same publication. The difficulty of replication in the absence of original vintage data sets was exemplified above for Woglom; another researcher would replicate an author's data set only by the coincidence of choosing the same issues of a publication as had been chosen by the original researcher.

We conducted a simulation experiment to study the effects of data revisions on the replication of a published empirical article. Our experiment is based on Goldberg and Saunders' three-equation model of the growth of agencies, branches, and subsidiaries of foreign banks in the United States. Our experiment simulates a researcher attempting to replicate Goldberg and Saunders' published results by collecting data from the *Survey of Current Business*. The complete experiment is comprised of 500 trials. Each trial consists of collecting a complete time-series on imports, investment,

and GNP from the *Survey* and estimating the Goldberg-Saunders' model. We constructed the data base for the experiment by collecting all preliminary and revised values for imports, investment, and GNP in each of the 40 quarters 1972:IV through 1982:III which had been published in the *Survey of Current Business* through the end of 1982. We used 118 monthly issues of the *Survey*.

Each trial of the experiment begins by drawing an integer in the range [1,118] from a uniform distribution, selecting the corresponding issue of the *Survey of Current Business*, and collecting all observations on imports, investment, and GNP from the issue. This process is repeated until the time-series for the variables are complete. Data once entered on the worksheet are never replaced so that a newer revised number does not displace an older preliminary or revised value already recorded. Each trial has the same number of observations as Goldberg and Saunders' data set. This procedure certainly is not the way Goldberg and Saunders collected their data. Nevertheless, our experiment illustrates the range and frequency distribution of estimates obtainable from randomly selected preliminary and revised data.

In the Goldberg-Saunders model, the values of assets held by agencies, branches, and subsidiaries of foreign banks in the United States are the dependent variables of the first, second, and third equations, respectively. Each equation contains the same independent variables: an intercept term; the lagged spread between yields on assets and liabilities; real gross private domestic investment; the ratio of imports to GNP; and a dummy variable to capture expectations of passage of the International Banking Act. We present below estimates of their model based on the most-recently published (newest) data, the first-ever-published (oldest) data, and the 500 trials discussed above.

In their article, Goldberg and Saunders report only that their model was estimated by generalized least squares; a replicator therefore could reasonably choose either a single-iteration or iterative Prais-Winston estimator. To study how this choice may affect the replication, we repeated our complete experiment for two estimation strategies. In the first, we estimated the agencies equation by ordinary least squares and the other two equations by a single-iteration Cochrane-Orcutt estimator, for comparability to Goldberg and Saunders. In the second, we estimated each equation by ordinary least squares; tested for first-order autocorrelation (Durbin-Watson statistics in the inconclusive region were considered significant); and, if necessary, estimated the equation by an iterative Prais-Winston estimator. We present detailed results only for the coefficient of gross private domestic investment. While this coefficient displayed the worst behavior, the coefficients of the other variables are roughly comparable. The frequency distributions of the coefficient estimates and t-statistics are shown in Figures 1–6 for the Goldberg-Saunders estimator and in Figures 7–12 for the iterative Prais-Winsten estimator.

The most significant finding of the replication experiment is the strikingly different coefficient estimates and significance levels obtained by use of the two different estimators. All 500 trials based on the Goldberg-Saunders estimation procedure produced significant t-statistics at the 5 percent level for the coefficient of investment in all three equations, similar to Goldberg and Saunders' results (see Figures 2, 4, and 6). In trials based on the Prais-Winsten estimator, one-half of t-statistics for the agencies equation (254 of 500), two-thirds of t-statistics for the branches equation (337 to 500), and one-third of t-statistics for the subsidiaries equation (151 of 500) are insignificant (see Figures 8, 10, and 12).

Our experiment demonstrates that Goldberg and Saunders' results are not easily replicated using data from published sources. A researcher using either the most-recently published data or a mixture of vintages of data from the *Survey of Current Business* would be unlikely to reproduce the Goldberg-Saunders findings and, in turn, may be misled regarding the value of Goldberg and Saunders' results as a foundation for future research.

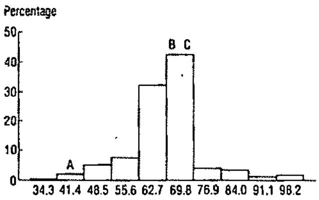


FIGURE 1. AGENCIES EQUATION

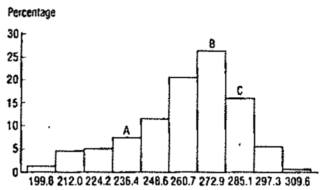


FIGURE 3. BRANCHES EQUATION

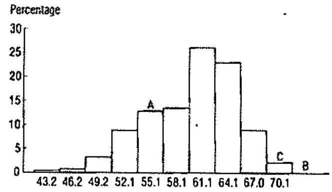


FIGURE 5. SUBSIDIARIES EQUATION

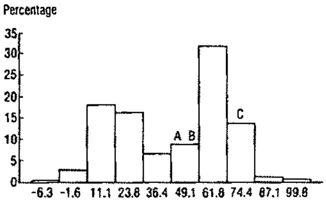


FIGURE 7. AGENCIES EQUATION

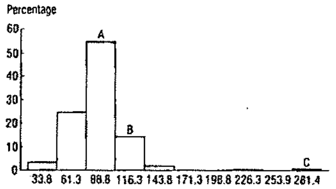


FIGURE 9. BRANCHES EQUATION

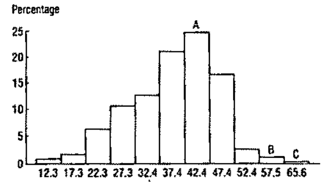


FIGURE 11. SUBSIDIARIES EQUATION

Note: Frequency Distribution of Investment Coefficient (Coefficient value, Midpoint of interval). Figure 1: Ordinary least squares estimator; Figures 3 and 5: Single-iteration Cochrane-Orcutt estimator; Figures 7, 9, and 11: Iterative Prais-Winsten estimator. A = estimate with most recently published data; B = estimate with originally published data; C = estimate in published article.

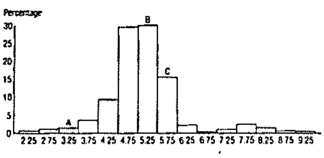


FIGURE 2. AGENCIES EQUATION
INVESTMENT COEFFICIENT

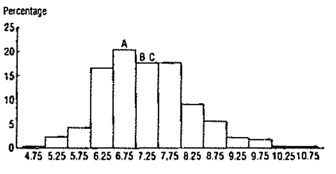


FIGURE 4. BRANCHES EQUATION
INVESTMENT COEFFICIENT

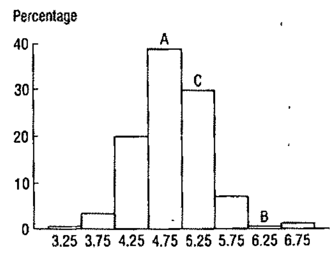


FIGURE 6. AGENCIES EQUATION
INVESTMENT COEFFICIENT

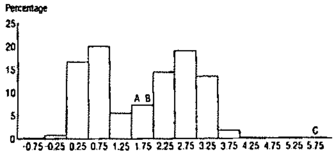


FIGURE 8. AGENCIES EQUATION
INVESTMENT COEFFICIENT

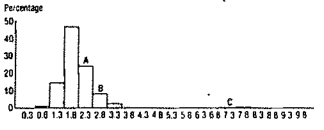


FIGURE 10. BRANCHES EQUATION
INVESTMENT COEFFICIENT

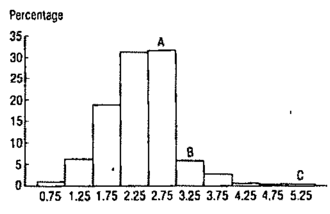


FIGURE 12. SUBSIDIARIES EQUATION
INVESTMENT COEFFICIENT

Note: Frequency Distribution of t-Ratio (Value of t-ratio, midpoint of interval). Figure 2: Ordinary least squares estimator; Figures 4 and 6: Single-iteration Cochrane-Orcutt estimator; Figures 8, 10, and 12: Iterative Prais-Winsten estimator. A, B, C are as defined above.

VI. Replication and Graduate Education in Economics

Several authors have suggested that replication of published empirical studies should be made an important part of graduate education in economics, thereby encouraging new professional economists to regard replication of past research as a legitimate starting point for new research; see, for example, the discussion in Edward Kane. The lack of collected data from published articles has frustrated attempts to implement this suggestion in the past. As a partial remedy, we have publicized the available *JMCB Project* data sets. In August 1983, we wrote to major economics departments with graduate degree programs, advertising the availability of the *JMCB Project* data sets as a basis for replication studies in graduate courses. Since 1983, students in advanced econometrics courses at Ohio State have been required to replicate and extend one published empirical study. Most students have chosen articles by authors who submitted data to the *JMCB Project*.

Overall, we received 59 requests for data between early 1983 and the conclusion of the *JMCB Project* in September 1984. Most requests were from other researchers, but many were from graduate students pursuing dissertation research. The research of Walter Kramer et al. (1985), for example, was greatly simplified by the availability of *JMCB Project* data sets.

VII. Conclusions and Recommendations

The replication of research is an essential component of scientific methodology. Only through replication of the results of others can scientists unify the disparate findings of various researchers in a discipline into a defensible, consistent, coherent body of knowledge.[5] While Ragner Frisch recognized this role of scientific replication in the first issue of *Econometrica*, today no major economics journal except the *JMCB* requests that authors submit programs and data sets. Referees of empirical articles must assume that the programming and data are correct, and readers of the articles find replication a difficult, frustrating, unrewarding, and often impossible task.

It is widely recognized that errors occur in empirical economic research and appear in published empirical articles. Our results from the *JMCB Project* suggest that such errors may be quite common. While many errors appear not to affect the conclusions of the authors significantly, the presence of the errors in a data set frustrates replication and prevents later researchers from building on earlier research. Some authors recognize that their research should build on earlier work but are forced by the unavailability of original data to employ *ad hoc* tests for the comparability of their data with those used in previous studies.[6]

Collinearity of data and high correlations among coefficient estimators are a widely recognized problem in studies based on time-series data. In these circumstances, even slight differences in data values or in the numerical precision of computer programs may produce sharply different parameter estimates. The existence of high collinearity increases both the difficulty of replication and the necessity for the preservation of the authors' original programs and data sets. Authors should appropriately investigate the numerical stability of their estimates and publish the results.

A reviewer of our original NSF funding application argued that the *JMCB Project* was unnecessary because "...all one had to

[5]The existence of a large number of studies which differ in their data sets and statistical estimators but nonetheless accept the same economic models and hypotheses also may provide a consistent body of scientific knowledge. New studies generally present an alternative view or interpretation of observed phenomena and attempt to extend results reached by previous researchers, not to confirm them. Replication is an essential element in the evaluation and unification of the results of any large group of studies.

[6]See, for example, Batavaia and Lash.

do was ask authors for their data." In principle, we agree with the reviewer and with the editorial statement of the *JPE* quoted above: authors of empirical articles should feel a professional responsibility to maintain program and data files for a reasonable amount of time after publication, and to provide these to other researchers. In fact, little benefit accrues to authors by providing programs and data, and little reward accrues to researchers conducting replication studies unless they can show that a major scientist has committed either fraud or a significant error in his research. This situation is not unique to economics, and similar issues regarding the availability of data are present in many fields; see, for example, James Craig and Sandra Reese (1973), Stephen Ceci and Elaine Walker (1983) and references therein.[7]

In this climate, replication may become viewed as a form of professional head-hunting rather than as an essential component of scientific research. Kane likens the process to petroleum exploration and mining. "Success" in replication may come to be defined as the discovery of error or fraud by another researcher, such that a publishable correction or comment arises, and a replicator may seek out those articles (geological sites) which have the highest likelihood of yielding success (oil). Kane's apt analogy illustrates why, in the present system, those authors whose research is the subject of a replication effort may interpret the very *act* of replication as a challenge to their professional competence and integrity. If programs and data were available from journals and replication became commonplace, authors would be less likely to feel threatened by replication, particularly if they have accurately recorded and carefully documented their programs and data.

One of the authors of this article was editor of the *JMCB* during most of the period when the articles included in the *JMCB Project* were published. It would be embarrassing to reveal the findings of the *Project* save for our belief that the findings would be little different if articles and authors were selected from any other major economics journal. In private correspondence, the editor of another major journal (not the *AER*) confided that he shares our belief.

On the basis of our findings, we recommend that journals require the submission of programs and data at the time empirical papers are submitted. The description of sources, data transformations, and econometic estimators should be so exact that another researcher could replicate the study and, it goes without saying, obtain the same results. This policy has three significant advantages.

First, authors would be able to supply programs and data sets at lower cost when their research is just completed than when a published article appears. We found that the frequency and magnitude of errors is smaller in data sets compiled by the researcher immediately after completion of the manuscript than in data sets compiled a year or more later. Furthermore, the compilation of programs and data by authors often uncovered ambiguities, errors, oversights, and misstatements which otherwise would have gone undiscovered.

Second, the journal provides a centralized cost-effective facility for distributing programs and data sets to other researchers. We recognize that journals will incur costs in handling and distributing these materials, costs which will be passed on to subscribers, authors, and those requesting the materials. Our experience at the *JMCB* is that the handling costs of such materials are low. During the last four years, authors have submitted more than 100 data sets on computer cards, magnetic tape, floppy disks, and paper printouts. These materials are stored in a portion of a single filing cabinet and are easily duplicated at low cost. Each requested data set is furnished at the actual cost of duplication including photocopying and computer charges. The organization and quality of submitted data sets is steadily improving, further reducing handling and distribution costs for the *JMCB*. Currently, an undergraduate economics student handles the receipt,

[7] We are indebted to Stephen Stigler for these references.

duplication, and mailing of requested data sets in approximately 2–4 hours per week.[8]

Finally, we believe these costs are more than offset by the benefits. The principal benefit is a reduced frequency of error in empirical articles due to more careful preparation of programs and data than might otherwise be the case. Subscribers benefit from the ability to obtain underlying programs and data. Authors benefit by publishing articles which can be replicated by other researchers, thereby increasing citations to the article and its importance as a basis for further research in the field. Editors and referees benefit by having access to authors' programs and data while evaluating papers for publication. Other researchers benefit by a substantial reduction in the costs of replicating published articles.

Alternatives must be proposed for authors whose research is based upon proprietary, licensed, or confidential programs and data sets such as SAS, SPSS, and Survey Research Center or *National Longitudinal Survey* data bases. Similar problems are rapidly appearing for researchers working on microcomputers, where software licensing agreements expressly prohibit copying or distribution of the software for any purpose. Authors should submit the version and serial numbers of proprietary programs (such as SAS, SPSS, RATS, and TSP) as well as listings of the instructions executed by the program. This audit trail allows replicators to trace bugs in the programs, changes in algorithms, and related difficulties. Users of large proprietary data sets should submit both the serial, version, or identification numbers and the date on which the data set was created or purchased. Although license agreements may restrict authors from furnishing data to other researchers, they should retain a copy of their data since vendors of data often are unable to reproduce the original vintage data set used by the researcher.

We emphasize that, in principle, similar rules should apply to simulation studies. This is a straightforward matter for small studies programmed in FORTRAN or using popular packaged programs. Our experience with the MPS model suggests, however, that formidible difficulties exist for studies based on large-scale models.

[8]No error checking or replication of data sets is currently performed by the *JMCB* or its staff.

REFERENCES

Batavia, Bala and Lash, Nicholas A., "The Impact of Bank Portfolio Composition on GNP," *Journal of Money, Credit and Banking*, November 1982, *14*, 517–24.

Broad, William and Wade, Nicholas, *Betrayers of Truth*, New York: Simon and Schuster, 1983.

Canarella, Giorgio and Garston, Neil, "Monetary and Public Debt Shocks: Tests and Efficient Estimates," *Journal of Money, Credit and Banking*, May 1983, *15*, 199–211.

Ceci, Stephen and Walker, Elaine, "Private Archives and Public Needs," *American Psychologist*, April 1983, *38*, 414–23.

Craig, James and Reese, Sandra, "Retention of Raw Data: A Problem Revisited," *American Psychology*, August 1973, *28*, 723.

Engle, Robert F., "Estimates of the Variance of U.S. Inflation Based Upon the ARCH Model," *Journal of Money, Credit and Banking*, August 1983, *15*, 286–301.

Feige, Edgar, "The Consequences of Journal Editorial Policies and a Suggestion for Revision," *Journal of Political Economy*, December 1975, *83*, 1291–95.

Feldstein, Martin, "Social Security, Induced Retirement, and Aggregate Capital Accumulation," *Journal of Political Economy*, September/October 1974, *82*, 905–26.

Friedman, Benjamin, "Federal Reserve Policy, Interest Rate Volatility, and the U.S. Capital Raising Mechanism," *Journal of Money, Credit and Banking*, November 1982, *14*, 721–45.

______ **and Johnson, David**, "Documentation for Simulation Exercises with the MPS Model as Modified to Incorporate Structural Models of the Corporate and Government Bond Markets," mimeo., Department of Economics, Harvard University, 1983.

Frisch, Ragnar, "Editorial," *Econometrica*, January 1933, *1*, 1–4.

Goldberg, Lawrence G. and Saunders, Anthony,

"The Growth of Organizational Forms of Foreign Banks in the U.S.," *Journal of Money, Credit and Banking*, August 1981, *13*, 365–74.

Gramlich, Edward M., "Models of Inflation Expectations Formation: A Comparison of Household and Economist Forecasts," *Journal of Money, Credit and Banking*, May 1983, *15*, 155–73.

Johannes, James M. and Rasche, Robert H., "Can the Reserves Approach to Monetary Control Really Work?," *Journal of Money, Credit and Banking*, August 1981, *13*, 298–313.

Kane, Edward J., "Why Journal Editors Should Encourage the Replication of Applied Econometric Research," *Quarterly Journal of Business and Economics*, Winter 1984, *23*, 3–8.

Kramer, Walter et al., "Diagnostic Checking in Practice," *Review of Economics and Statistics*, February 1985, *67*, 118–23.

Kuhn, Thomas S., *The Structure of Scientific Revolutions*, 2nd ed., Chicago: University of Chicago Press, 1970.

Leimer, Dean R. and Lesnoy, Selig D., "Social Security and Private Savings: New Time-Series Evidence," *Journal of Political Economy*, June 1982, *90*, 606–29.

Maris, Brian A., "Indirect Evidence on the Efficacy of Selective Credit Controls: The Case of Consumer Credit," *Journal of Money, Credit and Banking*, August 1981, *13*, 388–90.

Mayer, Thomas and Nathan, Harold, "Mortgage Rates and Regulation Q," *Journal of Money, Credit and Banking*, February 1983, *15*, 107–15.

______ **and** ______, "Errata," *Journal of Money, Credit and Banking*, May 1984, *16*, 271.

Merrick, John J., "Financial Market Efficiency, the Decomposition of 'Anticipated' versus 'Unanticipated' Money Growth, and Further Tests of the Relation Between Money and Real Output," *Journal of Money, Credit and Banking*, May 1983, *15*, 221–32.

Roley, V. Vance, "The Response of Short-Term Interest Rates to Weekly Money Announcements," *Journal of Money, Credit and Banking*, August 1983, *15*, 344–54.

Siskind, Frederick, "Minimum Wage Legislation in the United States: Comment," *Economic Inquiry*, January 1977, *15*, 135–38.

Welch, Finis, "Minimum Wage Legislation in the United States," *Economic Inquiry*, September 1974, *12*, 285–318.

______, "Minimum Wage Legislation in the United States: Reply," *Economic Inquiry*, January 1977, *15*, 139–42.

Woglom, Geoffrey, "A Reexamination of the Role of Stocks in the Consumption Function and the Transmission Mechanism," *Journal of Money, Credit and Banking*, May 1981, *13*, 215–20.

Editors, "Editorial Comment," *Journal of Political Economy*, December 1975, *83*, 1295–96.

Evaluating the Econometric Evaluations of Training Programs with Experimental Data

By ROBERT J. LALONDE*

This paper compares the effect on trainee earnings of an employment program that was run as a field experiment where participants were randomly assigned to treatment and control groups with the estimates that would have been produced by an econometrician. This comparison shows that many of the econometric procedures do not replicate the experimentally determined results, and it suggests that researchers should be aware of the potential for specification errors in other nonexperimental evaluations.

Econometricians intend their empirical studies to reproduce the results of experiments that use random assignment without incurring their costs. One way, then, to evaluate econometric methods is to compare them against experimentally determined results.

This paper undertakes such a comparison and suggests the means by which econometric analyses of employment and training programs may be evaluated. The paper compares the results from a field experiment, where individuals were randomly assigned to participate in a training program, against the array of estimates that an econometrician without experimental data might have produced. It examines the results likely to be reported by an econometrician using nonexperimental data and the most modern techniques, and following the recent prescriptions of Edward Leamer (1983) and David Hendry (1980), tests the extent to which the results are sensitive to alternative econometric specifications.[1] The goal is to appraise the likely ability of several econometric methods to accurately assess the economic benefits of employment and training programs.[2]

Section I describes the field experiment and presents simple estimates of the program effect using the experimental data. Sections II and III describe how econometricians evaluate employment and training programs, and compares the nonexperimental estimates using these methods to the experimental results presented in Section I. Section II presents one-step econometric estimates of the program's impact, while more complex two-step econometric estimates are presented in Section III. The re-

*Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, IL 60637. This paper uses public data files from the National Supported Work Demonstration. These data were provided by the Inter-University Consortium for Political and Social Research. I have benefited from discussions with Mariam Akin, Orley Ashenfelter, James Brown, David Card, Judith Gueron, John Papandreou, Robert Willig, and the participants of workshops at the universities of Chicago, Cornell, Iowa, Princeton, and MIT.

[1] These papers depict a more general crisis of confidence in empirical research. Leamer (1983) argues that any solution to this crisis must divert applied econometricians from "the traditional task of identifying unique inferences implied by a specific model to the task of determining the range of inferences generated by a range of models." Other examples of this literature are Leamer (1985), Leamer and Herman Leonard (1983), and Michael McAleer, Adrian Pagan, and Paul Volker (1985).

[2] Examples of nonexperimental program evaluations are Orley Ashenfelter (1978), Ashenfelter and David Card (1985), Laurie Bassi (1983a, b; 1984), Thomas Cooley, Thomas McGuire, and Edward Prescott (1979), Katherine Dickinson, Terry Johnson, and Richard West (1984), Nicholas Kiefer (1979a, b), and Charles Mallar (1978).

sults of this study are summarized in the final section.

I. The Experimental Estimates

The National Supported Work Demonstration (NSW) was a temporary employment program designed to help disadvantaged workers lacking basic job skills move into the labor market by giving them work experience and counseling in a sheltered environment. Unlike other federally sponsored employment and training programs, the NSW program assigned qualified applicants to training positions randomly. Those assigned to the treatment group received all the benefits of the NSW program, while those assigned to the control group were left to fend for themselves.[3]

During the mid-1970s, the Manpower Demonstration Research Corporation (MDRC) operated the NSW program in ten sites across the United States. The MDRC admitted into the program AFDC women, ex-drug addicts, ex-criminal offenders, and high school dropouts of both sexes.[4] For those assigned to the treatment group, the program guaranteed a job for 9 to 18 months, depending on the target group and site. The treatment group was divided into crews of three to five participants who worked together and met frequently with an NSW counselor to discuss grievances and performance. The NSW program paid the treatment group members for their work. The wage schedule offered the trainees lower wage rates than they would have received on a regular job, but allowed their earnings to increase for satisfactory performance and attendance. The trainees could stay on their supported work jobs until their terms in the program expired and they were forced to find regular employment.

Although these general guidelines were followed at each site, the agencies that operated the experiment at the local level provided the treatment group members with different work experiences. The type of work even varied within sites. For example, some of the trainees in Hartford worked at a gas station, while others worked at a printing shop.[5] In particular, male and female participants frequently performed different sorts of work. The female participants usually worked in service occupations, whereas the male participants tended to work in construction occupations. Consequently, the program costs varied across the sites and target groups. The program cost $9,100 per AFDC participant and approximately $6,800 for the other target groups' trainees.[6]

The MDRC collected earnings and demographic data from both the treatment and the control group members at the baseline (when MDRC randomly assigned the participants) and every nine months thereafter, conducting up to four post-baseline inter-

[3] Findings from the NSW are summarized in several reports and publications. For a quick summary of the program design and results, see Manpower Demonstration Research Corporation (1983). For more detailed discussions see Dickinson and Rebecca Maynard (1981); Peter Kemper, David Long, and Craig Thornton (1981); Stanley Masters and Maynard (1981); Maynard (1980); and Irving Piliavin and Rosemary Gartner (1981).

[4] The experimental sample included 6,616 treatment and control group members from Atlanta, Chicago, Hartford, Jersey City, Newark, New York, Oakland, Philadelphia, San Francisco, and Wisconsin. Qualified AFDC applicants were women who (*i*) had to be currently unemployed, (*ii*) had spent no more than 3 months in a job in the previous 6 months, (*iii*) had no children less than six years old, and (*iv*) had received AFDC payments for 30 of the previous 36 months. The admission requirements for the other participants differed slightly from those of the AFDC applicants. For a more detailed discussion of these prerequisities, see MDRC.

[5] Kemper and Long present a list of NSW projects and customers (1981, Table IV.4, pp. 65–66). The trainees produced goods and services for organizations in the public (42 percent of program hours), nonprofit (29 percent of program hours), and private sectors.

[6] The cost per training participant is the sum of program input costs, site overhead costs, central administrative costs, and child care costs minus the value of the program's output. These costs are in 1982 dollars. If the trainees' subsidized wages and fringe benefits are viewed as a transfer instead of a cost, the program costs per participant are $3,100 for the AFDC trainees and $2,700 for the other trainees. For a more detailed discussion of program costs and benefits, see Kemper, Long, and Thornton.

TABLE 1—THE SAMPLE MEANS AND STANDARD DEVIATIONS OF PRE-TRAINING EARNINGS AND OTHER CHARACTERISTICS FOR THE NSW AFDC AND MALE PARTICIPANTS

	Full National Supported Work Sample			
	AFDC Participants		Male Participants	
Variable	Treatments	Controls	Treatments	Controls
Age	33.37	33.63	24.49	23.99
	(7.43)	(7.18)	(6.58)	(6.54)
Years of School	10.30	10.27	10.17	10.17
	(1.92)	(2.00)	(1.75)	(1.76)
Proportion	.70	.69	.79	.80
High School Dropouts	(.46)	(.46)	(.41)	(.40)
Proportion Married	.02	.04	.14	.13
	(.15)	(.20)	(.35)	(.35)
Proportion Black	.84	.82	.76	.75
	(.37)	(.39)	(.43)	(.43)
Proportion Hispanic	.12	.13	.12	.14
	(.32)	(.33)	(.33)	(.35)
Real Earnings	$393	$395	1472	1558
1 year Before	(1,203)	(1,149)	(2656)	(2961)
Training	[43]	[41]	[58]	[63]
Real Earnings	$854	$894	2860	3030
2 years Before	(2,087)	(2,240)	(4729)	(5293)
Training	[74]	[79]	[104]	[113]
Hours Worked	90	92	278	274
1 year Before	(251)	(253)	(466)	(458)
Training	[9]	[9]	[10]	[10]
Hours Worked	186	188	458	469
2 years Before	(434)	(450)	(654)	(689)
Training	[15]	[16]	[14]	[15]
Month of Assignment	−12.26	−12.30	−16.08	−15.91
(Jan. 78 = 0)	(4.30)	(4.23)	(5.97)	(5.89)
Number of Observations	800	802	2083	2193

Note: The numbers shown in parentheses are the standard deviations and those in the square brackets are the standard errors.

views. Many participants failed to complete these interviews, and this sample attrition potentially biases the experimental results. Fortunately the largest source of attrition does not affect the integrity of the experimental design. Largely due to limited resources, the NSW administrators scheduled a 27th-month interview for only 65 percent of the participants and a 36th-month interview for only 24 percent of the non-AFDC participants. None of the AFDC participants were scheduled for a 36th-month interview, but the AFDC resurvey during the fall of 1979 interviewed 75 percent of these women anywhere from 27 to 44 months after the baseline. Since the trainee and control group members were randomly scheduled for all of these interviews, this source of attrition did not bias the experimental evaluation of the NSW program.

Naturally, the program administrators did not locate all of the participants scheduled for these interviews. The proportion of participants who failed to complete scheduled interviews varied across experimental group, time, and target group. While the response rates were statistically significantly higher for the treatment as opposed to the control group members, the differences in response rates were usually only a few percentage points. For the 27th-month interview, 72 percent of the treatments and 68 percent of the control group members completed interviews. The differences in response rates were

TABLE 2—ANNUAL EARNINGS OF NSW TREATMENTS, CONTROLS, AND EIGHT CANDIDATE COMPARISON GROUPS FROM THE *PSID* AND THE *CPS-SSA*

			Comparison Group[a,b]							
Year	Treatments	Controls	*PSID*-1	*PSID*-2	*PSID*-3	*PSID*-4	*CPS-SSA*-1	*CPS-SSA*-2	*CPS-SSA*-3	*CPS-SSA*-4
1975	$895	$877	7,303	2,327	937	6,654	7,788	3,748	4,575	2,049
	(81)	(90)	(317)	(286)	(189)	(428)	(63)	(250)	(135)	(333)
1976	$1,794	$646	7,442	2,697	665	6,770	8,547	4,774	3,800	2,036
	(99)	(63)	(327)	(317)	(157)	(463)	(65)	(302)	(128)	(337)
1977	$6,143	$1,518	7,983	3,219	891	7,213	8,562	4,851	5,277	2,844
	(140)	(112)	(335)	(376)	(229)	(484)	(68)	(317)	(153)	(450)
1978	$4,526	$2,885	8,146	3,636	1,631	7,564	8,518	5,343	5,665	3,700
	(270)	(244)	(339)	(421)	(381)	(480)	(72)	(365)	(166)	(593)
1979	$4,670	$3,819	8,016	3,569	1,602	7,482	8,023	5,343	5,782	3,733
	(226)	(208)	(334)	(381)	(334)	(462)	(73)	(371)	(170)	(543)
Number of Observations	600	585	595	173	118	255	11,132	241	1,594	87

[a]The Comparison Groups are defined as follows: *PSID*-1: All female household heads continuously from 1975 through 1979, who were between 20 and 55-years-old and did not classify themselves as retired in 1975; *PSID*-2: Selects from the *PSID*-1 group all women who received AFDC in 1975; *PSID*-3: Selects from the *PSID*-2 all women who were not working when surveyed in 1976; *PSID*-4: Selects from the *PSID*-1 group all women with children, none of whom are less than 5-years-old; *CPS-SSA* − 1: All females from Westat *CPS-SSA* sample; *CPS-SSA*-2: Selects from *CPS-SSA*-1 all females who received AFDC in 1975; *CPS-SSA*-3: Selects from *CPS-SSA*-1 all females who were not working in the spring of 1976; *CPS-SSA*-4: Selects from *CPS-SSA*-2 all females who were not working in the spring of 1976.

[b]All earnings are expressed in 1982 dollars. The numbers in parentheses are the standard errors. For the NSW treatments and controls, the number of observations refer only to 1975 and 1979. In the other years there are fewer observations, especially in 1978. At the time of the resurvey in 1979, treatments had been out of Supported Work for an average of 20 months.

larger across time and target group. For example, 79 percent of the scheduled participants completed the 9th-month interview, while 70 percent completed the 27th-month interview. The AFDC participants responded at consistently higher rates than the other target groups; 89 percent of the AFDC participants completed the 9th-month interview as opposed to 76 percent of the other participants. While these response rates indicate that the experimental results may be biased, especially for the non-AFDC participants, comparisons between the baseline characteristics of participants who did and did not complete a 27th-month interview suggest that whatever bias exists may be small.[7]

Table 1 presents some sample statistics describing the baseline characteristics of the AFDC treatment and control groups as well as those of the male NSW participants in the other three target groups.[8] As would be expected from random assignment, the

[7]This study evaluates the AFDC females separately from the non-AFDC males. This distinction is common in the literature, but it is also motivated by the differences between the response rates for the two groups. The Supported Work Evaluation Study (*Public Use Files User's Guide*, Documentation Series No. 1, pp. 18–27) presents a more detailed discussion of sample attrition. My working paper (1984, tables 1.1 and 2.3), compares the characteristics and employment history of the full NSW sample to the sample with pre- and postprogram earnings data. Randall Brown (1979) reports that there is no evidence that the response rates affect the experimental estimates for the AFDC women or ex-addicts, while the evidence for the ex-offenders and high school dropouts is less conclusive.

[8]The female participants from the non-AFDC target groups were not surveyed during the AFDC resurvey in the fall of 1979 and consequently do not report 1979 earnings and are not included with the AFDC sample. Excluding these women from the analysis does not affect the integrity of the experimental design.

TABLE 3—ANNUAL EARNINGS OF NSW MALE TREATMENTS, CONTROLS, AND SIX CANDIDATE COMPARISON GROUPS FROM THE *PSID* AND *CPS-SSA*

			Comparison Group[a,b]					
Year	Treatments	Controls	*PSID*-1	*PSID*-2	*PSID*-3	*CPS-SSA*-1	*CPS-SSA*-2	*CPS-SSA*-3
1975	$3,066	$3,027	19,056[a]	7,569	2,611	13,650	7,387	2,729
	(283)	(252)	(272)	(568)	(492)	(73)	(206)	(197)
1976	$4,035	$2,121	20,267	6,152	3,191	14,579	6,390	3,863
	(215)	(163)	(296)	(601)	(609)	(75)	(187)	(267)
1977	$6,335	$3,403	20,898	7,985	3,981	15,046	9,305	6,399
	(376)	(228)	(296)	(621)	(594)	(76)	(225)	(398)
1978	$5,976	$5,090	21,542	9,996	5,279	14,846	10,071	7,277
	(402)	(227)	(311)	(703)	(686)	(76)	(241)	(431)
Number of Observations	297	425	2,493	253	128	15,992	1,283	305

[a]The Comparison Groups are defined as follows: *PSID*-1: All male household heads continuously from 1975 through 1978, who were less than 55-years-old and did not classify themselves as retired in 1975; *PSID*-2: Selects from the *PSID*-1 group all men who were not working when surveyed in the spring of 1976; *PSID*-3: Selects from the *PSID*-1 group all men who were not working when surveyed in either spring of 1975 or 1976; *CPS-SSA*-1: All males based on Westat's criteria, except those over 55-years-old; *CPS-SSA*-2: Selects from *CPS-SSA*-1 all males who were not working when surveyed in March 1976; *CPS-SSA*-3: Selects from the *CPS-SSA*-1 unemployed males in 1976 whose income in 1975 was below the poverty level.

[b]All earnings are expressed in 1982 dollars. The numbers in parentheses are the standard errors. The number of observations refer only to 1975 and 1978. In the other years there are fewer observations. The sample of treatments is smaller than the sample of controls because treatments still in Supported Work as of January 1978 are excluded from the sample, and in the young high school target group there were by design more controls than treatments.

means of the characteristics and pretraining hours and earnings of the experimental groups are nearly the same. For example, the mean earnings of the AFDC treatments and the AFDC controls in the year before training differ by $2, the mean age of the two groups differ by 3 months, and the mean years of schooling are identical. None of the differences between the treatment's and control's characteristics, hours, and earnings are statistically significant.

The first two columns of Tables 2 and 3 present the annual earnings of the treatment and control group members.[9] The earnings of the experimental groups were the same in the pre-training year 1975, diverged during the employment program, and converged to some extent after the program ended. The post-training year was 1979 for the AFDC females and 1978 for the males.[10]

Columns 2 and 3 in the first row of Tables 4 and 5 show that both the unadjusted and regression-adjusted pre-training earnings of the two sets of treatment and control group members are essentially identical. Therefore, because of the NSW program's experimental design, the difference between the post-training earnings of the experimental groups is an unbiased estimator of the training effect, and the other estimators described in columns 5–10(11) are unbiased estimators as well. The estimates in column 4 indicate that the

[9]All earnings presented in this paper are in 1982 dollars. The NSW Public Use Files report earnings in experimental time, months from the baseline, and not calendar time. However, my working paper describes how to convert the experimental earnings data to the annual data reported in Tables 2 and 3.

[10]The number of NSW male treatment group members with complete pre- and postprogram earnings is much smaller than the full sample of treatments or the partial sample of control group members. This difference is largely explained by the two forms of sample attrition discussed earlier. In addition, however, (*i*) this paper excludes all males who were in Supported Work in January 1978, or entered the program before January 1976; (*ii*) in one of the sites, the administrators randomly assigned .4 instead of one-half of the qualified high school dropouts into the treatment group.

TABLE 4—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW AFDC PARTICIPANTS USING COMPARISON GROUPS FROM THE *PSID* AND THE *CPS-SSA*[a,b]

		NSW Treatment Earnings Less Comparison Group Earnings				Difference in Differences: Difference in Earnings Growth 1975–79 Treatments Less Comparisons		Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–79		Controlling for All Observed Variables and Pre-Training Earnings	
	Comparison Group Earnings	Pre-Training Year, 1975		Post-Training Year, 1979							
Name of Comparison Group[d]	Growth 1975–79 (1)	Unad-justed (2)	Ad-justed[c] (3)	Unad-justed (4)	Ad-justed[c] (5)	Without Age (6)	With Age (7)	Unad-justed (8)	Ad-justed[c] (9)	Without AFDC (10)	With AFDC (11)
	2,942	−17	−22	851	861	833	883	843	864	854	–
Controls	(220)	(122)	(122)	(307)	(306)	(323)	(323)	(308)	(306)	(312)	
PSID-1	713	−6,443	−4,882	−3,357	−2,143	3,097	2,657	1746	1,354	1664	2,097
	(210)	(326)	(336)	(403)	(425)	(317)	(333)	(357)	(380)	(409)	(491)
PSID-2	1,242	−1,467	−1,515	1,090	870	2,568	2,392	1,764	1,535	1,826	–
	(314)	(216)	(224)	(468)	(484)	(473)	(481)	(472)	(487)	(537)	
PSID-3	665	−77	−100	3,057	2,915	3,145	3,020	3,070	2,930	2,919	–
	(351)	(202)	(208)	(532)	(543)	(557)	(563)	(531)	(543)	(592)	
PSID-4	928	−5,694	−4,976	−2,822	−2,268	2,883	2,655	1,184	950	1,406	2,146
	(311)	(306)	(323)	(460)	(491)	(417)	(434)	(483)	(503)	(542)	(652)
CPS-SSA-1	233	−6,928	−5,813	−3,363	−2,650	3,578	3,501	1,214	1,127	536	1,041
	(64)	(272)	(309)	(320)	(365)	(280)	(282)	(272)	(309)	(349)	(503)
CPS-SSA-2	1,595	−2,888	−2,332	−683	−240	2,215	2,068	447	620	665	–
	(360)	(204)	(256)	(428)	(536)	(438)	(446)	(468)	(554)	(651)	
CPS-SSA-3	1,207	−3,715	−3,150	−1,122	−812	2,603	2,615	814	784	−99	1,246
	(166)	(226)	(325)	(311)	(452)	(307)	(328)	(305)	(429)	(481)	(720)
CPS-SSA-4	1,684	−1,189	−780	926	756	2,126	1,833	1,222	952	827	–
	(524)	(249)	(283)	(630)	(716)	(654)	(663)	(637)	(717)	(814)	

[a]The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1979. Based on the experimental data, an unbiased estimate of the impact of training presented in col. 4 is $851. The first three columns present the difference between each comparison group's 1975 and 1979 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

[b]Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

[c]The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

[d]See Table 2 for definitions of the comparison groups.

earnings of the AFDC females were $851 higher than they would have been without the NSW program, while the earnings of the male participants were $886 higher.[11] Moreover, the other columns show that the econometric procedure does not affect these estimates.

[11]It is commonly believed that the NSW program had little impact on the earnings of the male participants (see MDRC; A. P. Bernstein et al., 1985). My working paper discusses why this estimated impact differs from the results discussed elsewhere. The 1978 earnings data were largely collected during the 36th-month interview, where the difference between the male treatment and control group members' earnings averaged $175 per quarter.

II. Nonexperimental Estimates

In addition to providing researchers with a simple estimate of the impact of an employment program, MDRC's experimental data can also be used to evaluate several nonexperimental methods of program evaluation. This section puts aside the NSW control group and evaluates the NSW program using some of the econometric procedures found in studies of the employment and training programs administered under the MDTA, CETA, and JTPA.[12]

[12]These acronyms refer to the Manpower Development and Training Act–1962, the Comprehensive Em-

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE *PSID* AND THE *CPS-SSA*[a,b]

		NSW Treatment Earnings Less Comparison Group Earnings				Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons		Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78		
	Comparison Group Earnings	Pre-Training Year, 1975		Post-Training Year, 1978						Controlling for All Observed Variables and
Name of Comparison Group[d]	Growth 1975–78 (1)	Unadjusted (2)	Adjusted[c] (3)	Unadjusted (4)	Adjusted[c] (5)	Without Age (6)	With Age (7)	Unadjusted (8)	Adjusted[c] (9)	Pre-Training Earnings (10)
Controls	$2,063	$39	$−21	$886	$798	$847	$856	$897	$802	$662
	(325)	(383)	(378)	(476)	(472)	(560)	(558)	(467)	(467)	(506)
PSID-1	$2,043	−$15,997	−$7,624	−$15,578	−$8,067	$425	−$749	−$2,380	−$2,119	−$1,228
	(237)	(795)	(851)	(913)	(990)	(650)	(692)	(680)	(746)	(896)
PSID-2	$6,071	−$4,503	−$3,669	−$4,020	−$3,482	$484	−$650	−$1,364	−$1,694	−$792
	(637)	(608)	(757)	(781)	(935)	(738)	(850)	(729)	(878)	(1024)
PSID-3	($3,322	($455	$455	$697	−$509	$242	−$1,325	$629	−$552	$397
	(780)	(539)	(704)	(760)	(967)	(884)	(1078)	(757)	(967)	(1103)
CPS-SSA-1	$1,196	−$10,585	−$4,654	−$8,870	−$4,416	$1,714	$195	−$1,543	−$1,102	−$805
	(61)	(539)	(509)	(562)	(557)	(452)	(441)	(426)	(450)	(484)
CPS-SSA-2	$2,684	−$4,321	−$1,824	−$4,095	−$1,675	$226	−$488	−$1,850	−$782	−$319
	(229)	(450)	(535)	(537)	(672)	(539)	(530)	(497)	(621)	(761)
CPS-SSA-3	$4,548	$337	$878	−$1,300	$224	−$1,637	−$1,388	−$1,396	$17	$1,466
	(409)	(343)	(447)	(590)	(766)	(631)	(655)	(582)	(761)	(984)

[a] The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is $886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

[b] Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

[c] The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

[d] See Table 3 for definitions of the comparison groups.

The researchers who evaluated these federally sponsored programs devised both experimental and nonexperimental procedures to estimate the training effect, because they recognized that the difference between the trainees' pre- and post-training earnings was a poor estimate of the training effect. In a dynamic economy, the trainees' earnings may grow even without an effective program. The goal of these program evaluations is to estimate the earnings of the trainees had they not participated in the program. Researchers using experimental data take the earnings of the control group members to be an estimate of the trainees' earnings without the program. Without experimental data, researchers estimate the earnings of the trainees by using the regression-adjusted earnings of a comparison group drawn from the population. This adjustment takes into account that the observable characteristics of the trainees and the comparison group members differ, and their unobservable characteristics may differ as well.

Any nonexperimental evaluation of a training program must explicitly account for these differences in a model describing the observable determinants of earnings and the process by which the trainees are selected into the program. However, unlike in an experimental evaluation, the nonexperimental estimates of the training effect depend crucially on the way that the earnings and participation equations are specified. If the econometric model is specified correctly, the nonexperimental estimates should be the same (within sampling error) as the training effect generated from the experimental data, but if there is a significant difference between the nonexperimental and the experi-

ployment and Training Act–1973, and the Job Training Partnership Act–1982.

mental estimates, the econometric model is misspecified.[13]

The first step in a nonexperimental evaluation is to select a comparison group whose earnings can be compared to the earnings of the trainees. Tables 2 and 3 present the mean annual earnings of female and male comparison groups drawn from the *Panel Study of Income Dynamics (PSID)* and Westat's Matched *Current Population Survey–Social Security Administration File (CPS-SSA)*. These groups are characteristic of two types of comparison groups frequently used in the program evaluation literature. The *PSID*-1 and the *CPS-SSA*-1 groups are large, stratified random samples from populations of household heads and households, respectively.[14] The other, smaller, comparison groups are composed of individuals whose characteristics are consistent with some of the eligibility criteria used to admit applicants into the NSW program. For example, the *PSID*-3 and *CPS-SSA*-4 comparison groups in Table 2 include females from the *PSID* and the *CPS-SSA* who received AFDC payments in 1975, and were not employed in the spring of 1976. Tables 2 and 3 show that the NSW trainees and controls have earnings histories that are more similar to those of the smaller comparison groups, whose characteristics are similar to theirs, than those of the larger comparison groups.[15]

The second step in a nonexperimental evaluation is to specify a model of earnings and program participation to adjust for differences between the trainees and comparison group members. Equations (1) through (4) describe a conventional model of earnings and program participation that is typical of the kind econometric researchers use for this problem:

$$(1) \quad y_{it} = \delta D_i + \beta X_{it} + b_i + n_t + \varepsilon_{it}$$

$$(2) \quad \varepsilon_{it} - \rho \varepsilon_{it-1} = \nu_{it}$$

$$(3) \quad d_{is} = y_{is} + \gamma Z_{is} + \eta_{is}$$

$$(4) \quad D_i = 1 \quad \text{if } d_{is} > 0; \quad D_i = 0 \quad \text{if } d_{is} < 0.$$

In equation (1), earnings in each period are a function of a vector of individual characteristics, X_{it}, such as age, schooling, and race for individual i in time t; a dummy variable indicating whether the individual participated in training in period $s+1$, D_i; and an error with individual- and time-specific components and a serially correlated transitory disturbance. The transitory disturbance follows the first-order serial corre-

[13] Thomas Fraker, Maynard, and Lyle Nelson (1984) describe a similar study using the NSW AFDC and Young High School Dropouts. Instead of focusing the study on models of earnings and program participation, their study evaluates several strategies for choosing matched comparison groups. They use grouped Social Security earnings data when comparing the annual earnings of the NSW treatments to the earnings of each of the comparison groups.

[14] The *PSID* file including the poverty subsample selects only women and men who were household heads continuously from 1975 to 1979, and 1978, respectively. The *CPS-SSA* file matches the March 1976 *Current Population Survey* with Social Security earnings. Only individuals in the labor force in March 1976 with nominal income less than \$20,000 and household income less than \$30,000 are in this sample. In 1976, 2 percent of the females and 21 percent of the males had earnings at the Social Security maximum. In this paper, females younger than 20 or older than 55 and males older than 55 are excluded from the comparison groups.

[15] Not only are the pre-training earnings of the *PSID*-3 comparison group in Table 2 similar to the earnings of the NSW experimental groups, but the characteristics of these groups are similar as well. The mean age for the *PSID*-3 women is 40.95; the mean years of schooling is 10.31; the proportion of high school dropouts is 0.63; the proportion married is 0.01; the proportion black is 0.85; and the proportion Hispanic is 0.03. I experimented with matching the comparison groups even more closely to the pre-training characteristics of the experimental sample. However, these closely matched comparison groups are extremely small. For example there were 57 women from the *PSID* who received welfare payments in 1975, were not employed at the time of the survey in 1976, resided in a metropolitan area, and had only school-age children. The mean earnings of this group were \$1,137 in 1975; \$673 in 1976; \$743 in 1977; \$1,222 in 1978; and \$1,697 in 1979.

lation process described in equation (2). Equations (3) and (4) specify the participation decision: an individual participates in training and is admitted into the program in period $s+1$ if the latent variable d_{is} rises above zero. The participation equation is typically rationalized by the notion that the supply of individuals who decide to participate in training depends on the net benefit they expect to receive from participation and on the demand of the program administrators for training participants. The participation latent variable is typically a function of a vector of characteristics Z_{is}, current earnings y_{is}, and an error.

The estimators described in the column headings in Tables 4 and 5 (as well as many others in the literature) are based on econometric specifications that place different restrictions on the training model represented by equations (1)–(4) (although one common restriction assumes that the unobservables in the earnings and participation equations are uncorrelated). These estimates are consistent only insofar as their restrictions are consistent with the data. The restrictions can be tested provided the nonexperimental data base has sufficient information on the pre-training earnings and demographic characteristics of the trainees and comparison group members. An econometrician is unlikely to take seriously an estimate based on a model that failed one of these specification tests. Therefore, the results of such tests can often aid the researcher in choosing among alternative estimates. It follows, then, that simply checking whether the nonexperimental estimates replicate the experimental results and whether these estimates vary across different econometric procedures is not the only motivation for comparing experimental to nonexperimental methods. By making this comparison, we can also discover whether the nonexperimental data alone reliably indicate when an econometric model is misspecified and whether specification tests, which are supposed to ensure that the econometric model is consistent with the data, lead researchers to choose the "right" estimator.

In practice, the available data affect the composition of the comparison groups and the flexibility of the econometric specifications. For example, since there is only one year of pre-training earnings data, we cannot evaluate all of the econometric procedures that have been used in the literature, nor can we test all of the econometric specifications analyzed in this paper with the nonexperimental data alone.[16]

Nevertheless, several one-step estimators are evaluated in Tables 4 and 5, starting with the simple difference between the treatment and comparison group members' post-training earnings in column 4. Column 5 presents this earnings difference controlling for age, schooling, and race. This cross-sectional estimator is based on a model where these demographic variables are assumed to adequately control for differences between the earnings of the trainees and comparison group members. Column 6 presents the difference between the two nonexperimental groups' pre- and post-training earnings growth. This estimator allows for an unobserved individual fixed effect in the earnings equation and for the possibility that individuals with low values of this unobservable are more likely to participate in training. The cross-sectional estimator described in column 5 is now biased since the training dummy variable is correlated with the error in the earnings equation. Differencing the earnings equation removes the fixed effect, leaving[17]

$$(5)\qquad y_{it}-y_{is}=\delta D_i+\beta\cdot AGE_i + (\eta_t-\eta_s)+\varepsilon_{it}-\varepsilon_{is}.$$

[16] One limitation of the NSW Public Use File is that there is only one year of pre-experimental data available in calendar time as opposed to experimental time. Consequently, there are several nonexperimental procedures which require more than a year of pre-training earnings data that are not evaluated in this paper. If additional data were available, it is possible that these procedures would adequately control for differences between the NSW treatments and comparison group members and that the results of the specification tests would correctly guide an econometrician away from some of the estimates presented in this paper to the estimates based on these other procedures. See John Abowd (1983), Ashenfelter, Ashenfelter and Card, Bassi (1983b, 1984), and James Heckman and Richard Robb (1985).

[17] The other demographic variables, schooling and race, are constant over time.

The comparison group's earnings growth represents the earnings growth that the trainees would have experienced without the program. However, since the trainees may experience larger earnings growth than the comparison group members simply because they are usually younger, column 7 presents the difference between the earnings growth of the two groups controlling for age.

Column 8 presents the difference between the post-training earnings of the treatment and comparison group members, holding constant the level of pre-training earnings, while the estimator in column 9 controls both for pre-training earnings and the demographic variables. These estimators are consistent when the model of program participation stipulates that the trainees' pre-program earnings fell (see Table 1) because some of the training participants experienced some bad luck in the years prior to training. In this case, we would expect the trainees' earnings to grow even without the program.[18] The difference in differences estimator in columns 6 and 7 is now biased, since the training dummy variable is correlated with the transitory component of pre-training earnings in equation (5).[19] Finally, columns 10 and 11 report the estimates of the training effects controlling for all observed variables. Besides the variables described earlier, the additional regressors are employment status in 1976, AFDC status in 1975, marital status, residency in a metropolitan area with more than 100,000 persons, and number of children.

[18]Researchers have observed this dip in pre-training earnings for successive MDTA and CETA cohorts since 1964. See Ashenfelter (Table 1); Ashenfelter and Card (Table 1); Bassi (1983a, Table 4.1); and Kiefer (1979a, Table 4-1).

[19]This estimator is similar to one devised by Arthur Goldberger (1972) (or see G. S. Maddala, 1983) to evaluate the Head Start Program where participation in the program depended on a child's test score plus a random error. Similarly, participation in a training program can be thought of as a function of pre-training earnings and a random error. My working paper shows that this estimator is consistent as long as the unobservables in the earnings and participation equations are uncorrelated, and all of the observable variables in the model are used as regressors in the earnings equation.

Unlike the experimental estimates, the nonexperimental estimates are sensitive both to the composition of the comparison group and to the econometric procedure. For example, many of the estimates in column 9 of Table 4 replicate the experimental results, while other estimates are more than $1,000 larger than the experimental results. More specifically, the results for the female participants (Table 4) tend to be positive and larger than the experimental estimate, while for the male participants (Table 5), the estimates tend to be negative and smaller than the experimental impact.[20] Additionally, the nonexperimental procedures replicate the experimental results more closely when the nonexperimental data include pre-training earnings rather than cross-sectional data alone or when evaluating female rather than male participants.

The sensitivity of the nonexperimental estimates to different specifications of the econometric model is not in itself a cause for alarm. After all, few econometricians expect estimators based on misspecified models to replicate the results of experiments. Hence the considerable range of estimates is understandable given that inconsistent estimators are likely to yield inaccurate estimates. Before taking some of these estimates too seriously, many econometricians at a minimum would require that their estimators be based on econometric models that are consistent with the pre-training earnings data. Thus, if the regression-adjusted difference between the post-training earnings of the two groups is going to be a consistent estimator of the training effect, the regression-adjusted pre-training earnings of the two groups should be the same.

Based on this specification test, econometricians might reject the nonexperimental estimates in columns 4–7 of Table 4 in favor of the ones in columns 8–11. Few econometricians would report the training effect of $870 in column 5, even though this estimate differs from the experimental result

[20]The magnitude of these training effects is similar to the estimates reported in studies of the 1964 MDTA cohort, the 1969–70 MDTA cohort, and the 1976–77 CETA cohort. (See my working paper, Table I.1.)

by only $19. If the cross-sectional estimator properly controlled for differences between the trainees and comparison group members, we would not expect the difference between the regression adjusted pre-training earnings of the two groups to be $1,550, as reported in column 3. Likewise, econometricians might refrain from reporting the difference in differences estimates in columns 6 and 7, even though all these estimates are within two standard errors of $3,000. As noted earlier, this estimator is not consistent with the decline in the trainees' pre-training earnings.

This point can also be made with the estimates for the NSW male participants (Table 5). For example, all but one of the difference in differences estimates in column 6 are within one standard error of the experimental estimate. Yet for two reasons it is unlikely econometricians would report these estimates. First, as the results in column 7 suggest, since the trainees are younger their earnings might be expected to grow faster than the earnings of the comparison group members even without training. Second, as shown in Table 1, the pre-training earnings of the male participants fell in the period before training, suggesting that the trainees' earnings will grow even if the program is ineffective. Here again, econometricians might turn to the considerable range of estimates in columns 8–10.

The results of these specification tests suggest that an econometrician might report one of the estimates in columns 8–11. However, even without the experimental data, a researcher would find that the estimated training effect is still sensitive to the set of variables included in the earnings equation and to the composition of the comparison group. In Table 4, the estimates using the female household heads with school-age children (*PSID*-4) as a comparison group differ by more than $1,000. The largest estimate overstates the experimental result by $1,300, while the smallest estimate is within $100 of the experimental estimate. Likewise in column 11, we find that the same estimator with different comparison groups yields a set of estimates that vary by more than $1,000. The estimates for the male participants exhibit the same sensitivity to the choice of a comparison group and to the set of variables used as regressors in the earnings equation. However, the estimated standard errors associated with these training effects are larger than for the female estimates, making it more difficult to draw many conclusions from these results.

Without additional data it is difficult to see how a researcher would choose a training effect from among estimates. Moreover, the nonexperimental data base alone does not allow the econometrician to test whether these estimates are based on econometric models that adequately control for differences between the earnings of the trainees and comparison group members. In this case, comparisons between the experimental and nonexperimental estimates is the best specification test available.[21]

Specification tests that use pre-training earnings data are an appealing means to choose between alternative estimates, but these tests are not themselves always sufficient to identify unreliable estimators. This point becomes clear when we compare the estimates using the *PSID*-3 comparison group (as defined in Table 2) and those using the NSW control group. The characteristics of these two groups are nearly the same, as are their unadjusted and adjusted pre-training earnings. In each case the cross-sectional estimator in column 5 appears to be an unbiased estimate of the training effect. Moreover, both sets of estimates are unaffected by alternative econometric procedures. Thus both the experimental and nonexperimental estimates pass the same specification tests; nevertheless the nonexperimental estimate is approximately $2,100 larger than the experimental result. If a researcher did not know that one set of estimates was based on an experimental data set, it is hard to see how she or he would

[21]Ashenfelter, Ashenfelter and Card, and Bassi (1984) have noted in their studies using nonexperimental data that their results are sensitive to alternative econometric specifications and that there is evidence for male training participants that the econometric models are misspecified.

choose between two estimates where one training effect is roughly 3.5 times larger than the other.

III. Two-Step Estimates

The unobservables in the earnings equation were uncorrelated with those in the participation equation in all of the econometric models analyzed in the previous section. If, instead, the unobservables are correlated, none of the one-step least squares procedures are consistent estimators of the training effect. Individuals with high unobservables in their participation equation are more likely to participate in training. Yet if the unobservables in the earnings and participation equations are negatively correlated, these individuals are likely to have relatively low earnings, even after controlling for the observable variables in the model. Consequently, least squares underestimates the impact of training.

James Heckman (1978) proposes a two-step estimator that controls for the correlation between the unobservables by using the estimated conditional expectation of the earnings error as a regressor in the earnings equation. If the errors in the earnings and participation equations are jointly normally distributed, this conditional expectation is proportional to the conditional expectation of the error in the participation equation. Using the notation introduced in the last section, this relationship is expressed formally as

$$(6)\quad E(b_i+\varepsilon_{it}|Z_i,D_i)=\rho\sigma_\varepsilon\left[D_i\frac{\phi(\gamma Z_i)}{1-\Phi(\gamma Z_i)}-(1-D_i)\frac{\phi(\gamma Z_i)}{\Phi(\gamma Z_i)}\right]=rH_i,$$

where Z_i is a vector of observed variables, ρ is the correlation between the unobservables in the model, σ_ε^2 is the variance of the unobservables in the earnings equation, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the normal density and distribution functions. Therefore the earnings equation can be rewritten as

$$(7)\quad Y_{it}=\delta D_i+\beta X_{it}+rH_i+v_i^*,$$

where v_i^* is an orthogonal error by construction. To estimate the training effect, δ, the researcher first uses the coefficients from a probit estimate of the reduced-form participation equation to calculate the conditional expectation, H_i, for both the trainees and comparison group members,[22] and, second, uses this estimate, $\hat{H}_i$, as a regressor in the earnings equation. The training effect is then estimated by least squares.[23]

Table 6 presents estimates for the female and male training participants using the NSW controls, the *PSID*-1 and *CPS-SSA*-1 as comparison groups.[24] Unless some variables are excluded from the earnings equation, the training effect in this procedure is identified by the nonlinearity of the probit function. Hence, the rows of Table 6 allow us to evaluate the sensitivity of these estimates to different exclusion restrictions. The second column associated with each set of training effects presents the estimated participation coefficient. If the unobservables are uncorrelated, this estimate should not be significantly different from zero. Therefore, these estimates allow us to test whether this restriction on the correlation between the unobservables is consistent with the nonex-

[22] This is a choice-based sampling problem, since the probability of being in the nonexperimental data set is high for the NSW treatment group members and low for the comparison group members. The estimated probability of participation depends not only on the observed variables but on the numbers of trainees and comparison group members. Heckman and Richard Robb (1985) show that this procedure is robust to choice-based sampling. For an example of an application of this estimator in the evaluation literature, see Mallar.

[23] Since the estimated value of this conditional expectation is used as a regressor instead of the true value, the estimated standard errors associated with the least squares estimates are inconsistent and must be corrected. See Heckman (1978,: 1979); William Greene (1981); John Ham (1982); and Ham and Cheng Hsiao (1984).

[24] The two-step estimates using the smaller comparison groups were associated with large estimated standard errors.

TABLE 6—ESTIMATED TRAINING EFFECTS USING TWO-STAGE ESTIMATOR

Variables Excluded from the Earnings Equation, but Included in the Participation Equation	Comparison Group	NSW AFDC Females: Training Dummy	NSW AFDC Females: Estimate of Expectation	NSW Males: Training Dummy	NSW Males: Estimate of Expectation
		Heckman Correction for Program Participation Bias, Using Estimate of Conditional Expectation of Earnings Error as Regressor in Earnings Equation — Estimate of Coefficient for			
Marital Status, Residency in an SMSA, Employment Status in 1976, AFDC Status in 1975, Number of Children	*PSID*-1	1,129	−894	−1,333	−2,357
		(385)	(396)	(820)	(781)
	CPS-SSA-1	1,102	−606	−22	−1,437
		(323)	(480)	(584)	(449)
	NSW Controls	837	−18	899	−835
		(317)	(2376)	(840)	(2601)
Employment Status in 1976, AFDC Status in 1975, Number of Children	*PSID*-1	1,256	−823	–	–
		(405)	(410)		
	CPS-SSA-1	439	−979	–	–
		(333)	(481)		
	NSW Controls	–	–	–	–
Employment Status in 1976, Number of Children	*PSID*-1	1,564	−552	−1,161	−2,655
		(604)	(569)	(864)	(799)
	CPS-SSA-1	552	−902	13	−1,484
		(514)	(551)	(584)	(450)
	NSW Controls	851	147	889	−808
		(318)	(2385)	(841)	(2603)
No Exclusion Restrictions	*PSID*-1	1,747	−526	−667	−2,446
		(620)	(568)	(905)	(806)
	CPS-SSA-1	805	−908	213	−1,364
		(523)	(548)	(588)	(452)
	NSW Controls	861	284	889	−876
		(318)	(2385)	(840)	(2601)

Notes: The estimated training effects are in 1982 dollars. For the females, the experimental estimate of impact of the supported work program was \$851 with a standard error of \$317. The one-step estimates from col. 11 of Table 4 were \$2,097 with a standard error of \$491 using the *PSID*-1 as a comparison group, \$1,041 with a standard error of \$503 using the *CPS-SSA*-1 as a comparison group, and \$854 with a standard error of \$312 using the NSW controls as a comparison group. Estimates are missing for the case of three exclusions using the NSW controls since AFDC status in 1975 cannot be used as an instrument for the NSW females. For the males, the experimental estimate of impact of the supported work program was \$886 with a standard error of \$476. The one-step estimates from col. 10 of Table 5 were \$−1,228 with a standard error of \$896 using the *PSID*-1 as a comparison group, \$−805 with a standard error of \$484 using the *CPS-SSA*-1 as a comparison group, and \$662 with a standard error of \$506 using the NSW controls as a comparison group. Estimates are missing for the case of three exclusions for the NSW males as AFDC status is not used as an instrument in the analysis of the male trainees.

perimental data, and to examine whether this specification test leads econometricians to choose the "right" estimator.

The experimental estimates in Table 6 are consistent with MDRC's experimental design. All of these estimates are nearly identical to the experimental results presented in Tables 4 and 5. And furthermore, since the unobservables are uncorrelated by design, the estimated participation coefficients are never significantly different from zero.

Turning to the nonexperimental estimates we find that although the instruments used to identify the earnings equation have some effect on the results, generally these estimates are closer to the experimental estimates than are the one-step estimates (in column 11 of Tables 4 and 5). For the females, the difference between the two-step and one-step estimates are small relative to the estimated standard errors, and the estimates of the participation coefficient are only

marginally significantly different from zero. Interestingly, in one case when the *PSID*-1 sample is used as a comparison group, the estimated participation coefficient is significant (the *t*-statistic is 2.25) and the training effect of $1,129 is $968 closer to the experimental result than the one-step estimate. Additionally, this estimate is identical to the estimate using the *CPS-SSA*-1 comparison group, whereas the one-step estimates differed by $1,056. However, if an econometrician reported this training effect, she or he would have to argue that variables such as place of residence and prior AFDC status do not belong in the earnings equation. Otherwise, the econometrician is left to choose between a set of estimates that vary by as much as $1,308.

The two-step estimates are usually closer than the one-step estimates to the experimental results for the male trainees as well. One estimate, which used the *CPS-SSA*-1 sample as a comparison group, is within $600 of the experimental result, while the one-step estimate falls short by $1,695. The estimates of the participation coefficients are negative, although unlike these estimates for the females, they are always significantly different from zero. This finding is consistent with the example cited earlier in which individuals with high participation unobservables and low earnings unobservables were more likely to be in training. As predicted, the unrestricted estimates are larger than the one-step estimates. However, as with the results for the females, this procedure may leave econometricians with a considerable range ($1,546) of imprecise estimates; although, like the results for the females, there is no evidence that the results of the specification tests would lead econometricians to choose the "wrong" estimator.

IV. Conclusion

This study shows that many of the econometric procedures and comparison groups used to evaluate employment and training programs would not have yielded accurate or precise estimates of the impact of the National Supported Work Program. The econometric estimates often differ significantly from the experimental results. Moreover, even when the econometric estimates pass conventional specification tests, they still fail to replicate the experimentally determined results. Even though I was unable to evaluate all nonexperimental methods, this evidence suggests that policymakers should be aware that the available nonexperimental evaluations of employment and training programs may contain large and unknown biases resulting from specification errors.[25]

This study also yields several other findings that may help researchers evaluate other employment and training programs. First, the nonexperimental procedures produce estimates that are usually positive and larger than the experimental results for the female participants, and are negative and smaller than the experimental estimates for the male participants. Second, these econometric procedures are more likely to replicate the experimental results in the case of female rather than male participants. Third, longitudinal data reduces the potential for specification errors relative to the cross-sectional data. Finally, the two-step procedure certainly does no worse than, and may reduce the potential for specification errors relative to, the one-step procedures discussed in Section II.

More generally, this paper presents an alternative approach to the sensitivity analyses proposed by Leamer (1983, 1985) and others for bounding the specification errors associated with the evaluation of economic hypotheses. This objective is accomplished by comparing econometric estimates with experimentally determined results. The data from an experiment yield simple estimates of the impact of economic treatments that are independent of any model specification. Successful econometric methods are intended to

[25] There is some evidence that this message has been passed on to the appropriate policymakers. See Recommendations of the Job Training Longitudinal Survey Research Advisory Panel to Office of Strategic Planning and Policy Development, U.S. Department of Labor, November 1985. This has led to at least a tentative decision to operate some part of the Job Training Partnership Act program sites using random assignment. (See Ernst Stromsdorfer et al., 1985.)

reproduce these estimates. The only way we will know whether these econometric methods are successful is by making the comparison. This paper takes the first step along this path, but there are other experimental data bases available to econometricians and much work remains to be done. For example, there have been several other employment and training experiments testing the effect of training on earnings, four Negative Income Tax Experiments testing hypotheses about labor supply, a medical insurance experiment testing hypotheses about insurance and medical demand, a housing experiment testing hypotheses about housing demand and supply, and a time-of-day electricity pricing experiment testing hypotheses about electricity demand.[26] There clearly remain many opportunities to use the experimental method to assess the potential for specification bias in the evaluation of social programs, and in other areas of econometric research as well.

[26]See Linda Aiken and Barbara Kehrer (1985), Abt Associates (1984), Gary Burtless (1985), Barbara Goldman (1981), Goldman et al. (1985), Jerry Hausman and David Wise (1985), J. Ohls and G. Carcagno (1978), and SRI International (1983).

REFERENCES

Abowd, John, "Program Evaluation," Working Paper, University of Chicago, 1983.

Aiken, Linda and Kehrer, Barbara, *Evaluation Studies Review Annual*, Vol. 10, Beverly Hills: Sage Publications, 1985.

Ashenfelter, Orley, "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, February 1978, *60*, 47–57.

______ **and Card, David,** "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, November 1985, *67*, 648–60.

Bernstein, A. P. et al., "The Forgotten Americans," *Business Week*, September 2, 1985, 50–55.

Bassi, Laurie, (1983a) "Estimating the Effect of Training Programs With Non-Random Selection," Princeton University, 1983.

______, (1983b) "The Effect of CETA on the Post-Program Earnings of Participants," *Journal of Human Resources*, Fall 1983, *18*, 539–556.

______, "Estimating the Effects of Training Programs with Nonrandom Selection," *Review of Economics and Statistics*, February 1984 *66*, 36–43.

Brown, Randall, "Assessing the Effects of Interview Nonresponse on Estimates of the Impact of Supported Work," Mathematica Policy Research Inc., Princeton, 1979.

Burtless, Gary, "Are Targeted Wage Subsidies Harmful? Evidence from a Wage Voucher Experiment," *Industrial and Labor Relations Review*, October 1985, *39*, 105–114.

Cooley, Thomas, McGuire, Thomas and Prescott, Edward, "Earnings and Employment Dynamics of Manpower Trainees: An Exploratory Econometric Analysis," in Ronald Ehrenberg, ed., *Research in Labor Economics*, Vol. 4, Suppl. 2, 1979, 119–47.

Dickinson, Katherine and Maynard, Rebecca, *The Impact of Supported Work on Ex-Addicts*, New York: Manpower Demonstration Research Corporation, 1981.

______, **Johnson, Terry and West, Richard,** *An Analysis of the Impact of CETA Programs on Participants' Earnings*, Washington: Department of Labor, Employment and Training Administration, 1984.

Fraker, Thomas, Maynard, Rebecca and Nelson, Lyle, *An Assessment of Alternative Comparison Group Methodologies for Evaluating Employment and Training Programs*, Princeton: Mathematica Policy Research Inc., 1984.

Goldberger, Arthur, "Selection Bias in Evaluating Treatment Effects," Discussion Paper No. 123–72, Institute for Research on Poverty, University of Wisconsin, 1972.

Goldman, Barbara, "The Impacts of the Immediate Job Search Assistance Experiment," Manpower Demonstration Research Corporation, New York, 1981.

______ **et al.,** "Findings From the San Diego Job Search and Work Experience Demonstration," New York: Manpower Demonstration Research Corporation, 1985.

Greene, William, "Sample Selection Bias as a Specification Error: Comment," *Econometrica*, May 1981, *49*, 795–98.

Ham, John, "Estimation of a Labor Supply Model with Censoring Due to Unemployment and Underemployment," *Review of Economic Studies*, July 1982, *49*, 335–54.

______ **and Hsiao, Cheng,** "Two-Stage Estimation of Structural Labor Supply Parameters Using Interval Data From the 1971 Canadian Census," *Journal of Econometrics*, January/February 1984, *24*, 133–58.

Hausman, Jerry A. and Wise, David A., *Social Experimentation*, NBER, Chicago: University of Chicago Press, 1985.

Heckman, James, "Dummy Endogenous Variables in a Simultaneous Equations System," *Econometrica*, July 1978, *46*, 931–59.

______, "Sample Selection Bias as a Specification Error," *Econometrica*, January 1979, *47*, 153–61.

______ **and Robb, Richard,** "Alternative Methods for Evaluating the Impact of Interventions: An Overview," Working Paper, University of Chicago, 1985.

Hendry, David, "Econometrices: Alchemy or Science?" *Economica*, November 1980, *47*, 387–406.

Kemper, Peter and Long, David, "The Supported Work Evaluation: Technical Report on the Value of In-Program Output Costs," Manpower Demonstration Research Corporation, New York, 1981.

______, ______, **and Thornton, Craig,** "The Supported Work Evaluation: Final Benefit-Cost Analysis," Manpower Demonstration Research Corporation, New York, 1981.

Kiefer, Nicholas, (1979a) *The Economic Benefits of Four Employment and Training Programs*, New York: Garland Publishing, 1979.

______, (1979b) "Population Heterogeneity and Inference from Panel Data on the Effects of Vocational Training," *Journal of Political Economy* October 1979, *87*, S213–26.

LaLonde, Robert, "Evaluating the Econometric Evaluations of Training Programs With Experimental Data," Industrial Relations Section, Working Paper No. 183, Princeton University, 1984.

Leamer, Edward, "Let's Take the Con Out of Econometrics," *American Economic Review*, March 1983, *73*, 31–43.

______, "Sensitivity Analysis Would Help," *American Economic Review*, June 1985, *75*, 308–13.

______ **and Leonard, Herman,** "Reporting the Fragility of Regression Estimates," *Review of Economics and Statistics*, May 1983, *65*, 306–12.

McAleer, Michael, Pagan, Adrian and Volker, Paul, "What Will Take the Con Out of Econometrics?," *American Economic Review*, June 1985, *75*, 293–306.

Maddala, G. S., *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press, 1983.

Mallar, Charles, "Alternative Econometric Procedures for Program Evaluations: Illustrations From an Evaluation of Job Corps Book," *Proceedings of the American Statistical Association*, 1978, 317–21.

______, **Kerachsky, Stuart and Thornton, Craig,** *The Short-Term Economic Impact of the Jobs-Corps Program*, Princeton: Mathematica Policy Research Inc., 1978.

Masters, Stanley and Maynard, Rebecca, "The Impact of Supported Work on Long-Term Recipients of AFDC Benefits," Manpower Demonstration Research Corporation, New York 1981.

Maynard, Rebecca, "The Impact of Supported Work on Young School Dropouts," Manpower Demonstration Research Corporation, New York, 1980.

Ohls, J. and Carcagno, G., *Second Evaluation of the Private Employment Agency Job Counsellor Project*, Princeton: Mathematica Policy Research Inc., 1978.

Piliavin, Irving and Gartner, Rosemary, "The Impact of Supported Work on Ex-Offenders," Manpower Demonstration Research Corporation, New York, 1981.

Stromsdorfer, Ernst et al., "Recommendations of the Job Training Longitudinal Survey Research Advisory Panel to the Office of Strategic Planning and Policy Development, U.S. Department of Labor," unpublished report, Washington, November 1985.

Abt Associates, "AFDC Homemaker-Home

Health Aid Demonstration Evaluation," 2nd Annual Report, Washington, 1984.

Manpower Demonstration Research Corporation, *Summary and Findings of the National Supported Work Demonstration*, Cambridge: Ballinger, 1983.

SRI International, *Final Report of the Seattle-Denver Income Maintenance Experiment: Design and Results*, Washington: Department of Health and Human Services, 1983.

A Test for Speculative Bubbles in the Sterling-Dollar Exchange Rate: 1981–84

By GEORGE W. EVANS*

The U.S. dollar price of the U.K. pound sterling is tested for a speculative bubble, defined as a period with a nonzero median in excess returns. A nonparametric procedure is developed which controls for data mining over the period of flexible exchange rates and finds a negative bubble in the excess return to holding sterling rather than dollar assets during 1981–84. Possible interpretations are bootstrap equilibria (rational bubbles), asymmetric fundamentals, and nonrational expectations.

Many observers have felt that the foreign exchange markets have been too volatile since the advent of flexible rates, subject to large and persistent speculative movements in exchange rates over substantial periods of time.[1] A salient example of this is the sustained decrease of the U.S. dollar price of the U.K. pound sterling from December 1980 through February 1985, a decline which, toward the latter part of the period, led to a belief among many economists and policymakers in both countries that the pound was undervalued relative to the dollar.[2]

Figure 1 indicates the magnitude of the swing: during this four-year period, the dollar price of the pound fell by over one-half. The extent of this slide cannot be explained by differential interest rates or inflation rates between the two countries. For example, correcting for interest rate differentials, the average excess return to holding dollars rather than sterling over this period was over 19 percent per annum.

The size and persistence of the negative excess returns to holding sterling might appear to constitute decisive evidence of a "speculative bubble" or "abnormality" in the dollar-pound exchange rate during this period. This conclusion, however, requires

*Department of Economics, Stanford University, Stanford, CA 94305-1992. This research was begun while I was visiting the London School of Economics, and was supported by the Center for Economic Policy Research, Stanford. I am indebted to Pauline Andrews for developing the computer programs required for the tests, to Simona Hughes for collecting the data, and to Roger Alford for providing the DIP program used to plot Figure 1. I also thank the participants in the Stanford Macroeconomics Seminar and the managing editor of this journal for their comments. An earlier version of this paper was circulated under the title "Speculative Bubbles and the Sterling-Dollar Exchange Rate: A New Test."

[1]See, for example, Ronald McKinnon (1979, pp. 155–56). There was a related discussion, antedating the 1971–73 shift to flexible exchange rates, over destabilizing and stabilizing speculation under freely floating exchange rates. References can be found in McKinnon.

[2]McKinnon (1984, ch. 5) argued in March 1984 that the dollar was overvalued and the *Economist* (January 5, 1985, p. 54) reported that an OECD study based on purchasing power parity calculations showed the pound substantially undervalued against the dollar in mid-1984. In the United States, the Treasury in 1983 described coordinated "intervention to stem the rise of the dollar" (*Wall Street Journal*, August 2, 1983). In July 1984, Paul Volcker, the Chairman of the Federal Reserve Board, discussed the risk of a fall in the dollar (*Financial Times*, July 26, 1984) and in early 1985 referred to an overvalued dollar, the need for concerted intervention, and the possibility of a "very sharp decline" in the dollar (*Economist*, March 16, 1985, p. 73 and *Financial Times*, March 7, 1985). In the United Kingdom, interest rates were increased in July 1984 "in the face of a further speculative run against the pound" (*Financial Times*, July 7, 9, and 12, 1984). Then in January 1985 the British government, which was "extremely concerned about the recent slide in sterling on the foreign exchanges," engineered a total four and one-half percentage point increase in bank base rates in just over two weeks in an effort to arrest the decline in the pound sterling (*Financial Times*, January 14, 15, and 29, 1985). Intervention in the foreign exchange markets by the central banks of the G5 countries occurred on a number of occasions (for example, see *Financial Times*, July 4, 1984, and March 22, 1985).

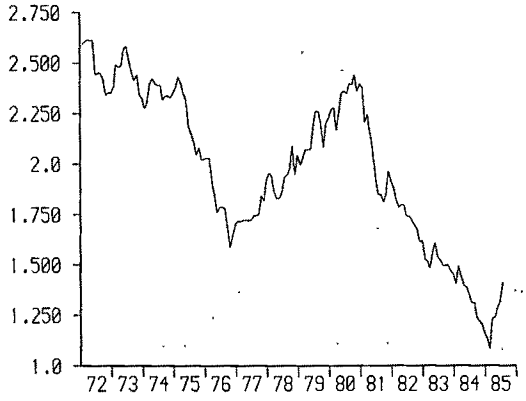


FIGURE 1. U.S. DOLLAR PRICE OF U.K. POUND STERLING

careful statistical support since, for example, an asset price following a random walk will on occasion, purely by chance, exhibit such apparent abnormalities over subperiods. A sound statistical test must therefore take account of data mining. The focus of this paper is the design of a procedure for detecting a precisely defined type of speculative bubble or abnormality in the dollar price of sterling over 1981–84, which protects against data mining from within the flexible exchange rate period.

In apparent contrast to the view that the foreign exchange markets contain periods of bubbles or abnormalities in excess returns, it has been widely assumed in the literature that, under flexible rates, the foreign exchange markets set prices efficiently. If agents are risk neutral, expectations are formed rationally, transactions costs can be neglected and markets are competitive, then "simple efficiency" holds and there should be no expected profit opportunities in the forward exchange markets. This is usually expressed as

$$(1) \qquad E_t x_{t+1} = 0 \quad \text{for} \quad x_{t+1} = s_{t+1} - f_t,$$

where s_{t+1} is the logarithm of the spot price of foreign currency in period $t+1$ and f_t is the logarithm of the one-period forward rate, that is, the price at time t of forward currency to be delivered at $t+1$. Here E_t denotes the mathematical expectation conditioned on information available at time t. Equation (1) is also referred to as the "unbiased expectations" hypothesis since it is equivalent to $f_t = E_t s_{t+1}$.

Equation (1) can be tested empirically in several ways, for example, by regressing x_t, the excess rate of return to holding a forward exchange contract, on variables in the information set and testing the null hypothesis of zero coefficients, or by testing the implied nonlinear restrictions imposed on vector autoregressive models. Although some tests have appeared to support simple efficiency, recent tests using larger sample sizes seem to have clearly detected small but statistically significant deviations from (1).[3] However, the dominant interpretation of these findings is that they are due to time-varying risk premia. Thus it is argued that $E_t \tilde{x}_{t+1} = 0$ for a suitably defined risk-adjusted excess return $\tilde{x}_{t+1}$, so that efficiency holds when allowance is made for risk-averse agents. We return to this issue below in Sections I, Part B, and II. Assuming that (1) holds, or a suitable modification of (1) to allow for risk premia,[4] can this be reconciled with the view of a market subject to large speculative movements? There are several possibilities.

Since to a close approximation the markets obey covered interest arbitrage, we have $f_t = s_t + i_{t,1} - i^*_{t,1}$ where $i_{t,1}$ and $i^*_{t,1}$ are the domestic and foreign one-period interest rates from period t to $t+1$, known at time t, on comparable assets such as Eurocurrency deposits. Then x_{t+1} is also given by

$$(2) \qquad x_{t+1} = s_{t+1} - s_t + i^*_{t,1} - i_{t,1},$$

[3]A partial list of the large and expanding literature on tests of simple efficiency includes Jacob Frenkel (1977, 1981), John Geweke and Edgar Feige (1979), Lars Peter Hansen and Robert Hodrick (1980), Craig Hakkio (1981), Sebastian Edwards (1983), and Richard Baillie, Robert Lippens and Patrick McMahon (1983), as well as several of the references cited later in this paper.

[4]To simplify discussion of the other issues involved, I ignore risk premia throughout the remainder of the introduction.

and (1) may be equivalently written as

$$(3) \qquad i_{t,1} = i^*_{t,1} + E_t s_{t+1} - s_t,$$

which is sometimes known as the open interest parity condition. Note that x_{t+1} can be interpreted as the excess rate of return, in domestic currency, from holding an uncovered foreign interest-bearing asset rather than a comparable domestic asset. It was shown in Rudiger Dornbusch (1976) that when (3) is incorporated in models with predetermined goods prices, it is possible for the exchange rate, in response to new information about fundamentals, to overshoot its long-run equilibrium value and thence to follow a systematic path back toward equilibrium. However, the predictable component of the movement in s_t would be entirely captured in interest rate differentials, which we net out when computing excess returns.

Another possible explanation of large systematic variations in exchange rates is that the rate follows a solution to (3) other than the one dictated by fundamentals.[5] The existence of a wide variety of solutions for asset or goods prices satisfying a condition analogous to (3) has been emphasized by John Taylor (1977), Olivier Blanchard (1981), Robert Flood and Peter Garber (1980), Blanchard and Mark Watson (1982), and Behzad Diba and Herschel Grossman (1983). Solutions other than the "market fundamentals" solution have been variously called "rational bubble," "sunspot," and "bootstrap" equilibria.[6] I will use the name "bootstrap" for such solutions and reserve the term "bubble" for a related but distinct phenomenon. The possibility of bootstrap solutions in the foreign exchange markets was explicitly noted by Flood and Garber (1982) using a simple complete model incorporating a continuous time version of (3), and I give examples in Section III below.

Indirect tests for bootstrap equilibria have been of three types.[7]

(*i*) Variance bounds tests of Robert Shiller (1981) and Blanchard and Watson which, when applied to equities, find too much volatility in equity prices given the behavior of dividends. The principal problem with this type of test is that there are several alternative explanations, discussed by Stephen LeRoy (1984), which could explain such excess volatility without the necessity of assuming prices deviating from their fundamental values. However, Kenneth West (1985a) has recently argued that at least some of these problems can be overcome and has provided additional evidence of excess volatility.

(*ii*) Specification tests, for example, West (1985b), which compare two estimates of parameters required to calculate the "fundamentals" solution. Like (*i*), this test requires an assumed model for pricing the asset.

(*iii*) Runs tests, applied by Blanchard and Watson to gold prices. Here the tests failed to detect an excess of long runs which would have suggested bubbles. However, the tests may have had low power for reasons discussed by Blanchard and Watson,[8] and for other reasons mentioned below. It should also be noted that a tendency toward runs can result from certain distributions of fundamentals.

This paper proposes a new test for speculative bubbles, closer in spirit to the runs test, and applies it to the exchange rate between the U.K. pound sterling and the U.S. dollar. In order to clarify what null hypothesis is being tested, I provide a precise definition of the type of abnormality in excess returns which will be referred to as a

[5] Which variables constitute fundamentals depends upon the specification of the complete model.

[6] The definition of the fundamentals solution is discussed in Edwin Burmeister, Flood, and Garber (1983) and Bennett McCallum (1983). See also Section III below. The entire class of solutions to such models is given in C. Gourieroux, J. J. Laffont, and A. Monfort (1982). The disequilibrium stability of bootstrap equilibria has been considered in my paper (1985).

[7] Direct tests have been conducted by Flood and Garber (1980) and Diba and Grossman (1984), with negative results.

[8] Suppose there is a bubble caused by a bootstrap solution which has probability π of continuing. For π near one or one-half, the tendency toward long runs becomes negligible.

speculative bubble. For the purposes of this paper, I define a speculative bubble as a subperiod during which there is a nonzero *median* in the distribution of x_t, the excess return to holding foreign currency (if an allowance is made for risk premia, I alter the definition to require a nonzero median in the distribution of the risk-adjusted excess returns, $\tilde{x}_t$).

This definition has a number of advantages. First, there is some basis in common usage, since it implies a tendency towards an abnormal number of positive or negative excess returns during the subperiod. Second, if efficiency holds, then the existence of a speculative bubble must correspond to a skewed distribution of excess returns. The low probability of large excess returns of opposite sign to those that predominate can be interpreted as the possibility of a crash. Finally, this definition of speculative bubble is well suited to testing, since it is stated in terms of one measure of the central tendency of excess returns.

It must be emphasized at the outset, however, that detection of a speculative bubble so defined does not imply the detection of a bootstrap equilibrium or "rational bubble" in the sense of Flood and Garber (1980) or Blanchard and Watson. Other possible causes of a speculative bubble are a skewed distribution of fundamental innovations and nonrational solutions.[9]

In the next section I describe and carry out an appropriate nonparametric test which finds evidence of a speculative bubble in the sterling-dollar exchange rate over 1981–84. The following section shows that the evidence for a bubble remains strong when an allowance is made for risk premia. The remaining sections discuss at length the possible interpretations of the results.

I. A New Test for Bubbles

One of the major difficulties in testing for speculative bubbles is that there may be only one or two bubbles covering only part of the sample period. Furthermore, if there are two or more bubbles, they may be of opposite signs. We cannot, therefore, assume a uniform pattern over the whole sample period, and a powerful test must be designed to reflect this.

A. *Description of the Test*

Let x_t be the stochastic process of excess returns to holding foreign currency for $t = 1,\dots,T$, and let m_t be the median of x_t in period t conditional on information through $t-1$. The null hypothesis H_0 is that $m_t = 0$ for $t = 1,\dots,T$, and we desire a test designed to detect alternatives H_1 under which $m_t = m \neq 0$ for $t = T_1,\dots,T_2$ where $1 \leq T_1 < T_2 \leq T$.

A suitable basis for a test of H_0 vs. H_1 is the sign test, a nonparametric test which looks at the difference between the number of positive and negative signs for x_t over the period.[10] This test would be directly applicable if the alternative were uniform, that is, $m_t = m \neq 0$ for $t = 1,\dots,T$. The sign test has the advantages that it explicitly focuses on the conditional median and that no assumptions are required about the distribution of x_t other than continuity. In contrast, the usual t-test is clearly not an appropriate starting point, since it is based on the sample mean and since the validity of the test in small samples depends on assumptions of constant variance and normal kurtosis, whereas it is widely believed that changes in asset prices typically are heteroscedastic and leptokurtic.

The statistical problem is to adapt the sign test so that it is sensitive to a departure from 0 median over only part of the sample period. Although the sign test applied to the full-sample period might detect such a deviation, this test is likely to have very low power for such alternatives. On the other hand, the application of the sign test to a suspect subperiod, ignoring data from the rest of the full period, will be an invalid test as a result of data mining. For my particular applica-

[9]Those readers who object to this use of "speculative bubble" may substitute for it another phrase such as "empirical bubble" or "abnormality in excess returns," or choose their own terminology.

[10]A description of the sign test can be found in W. J. Dixon and A. M. Mood (1946).

tion, the question is whether the behavior of the exchange rate over 1981–84 was too extreme to have plausibly occurred by chance, under the null hypothesis, controlling for data mining from the post-1972 period of flexible exchange rates.

There is no simple procedure for obtaining a properly adjusted significance level for the apparent negative median of x_t over 1981–84. At the time of writing, the pound sterling has floated against the dollar for twelve years and eleven months,[11] giving a full sample period of 155 months.[12] There are thus just over three independent four-year subperiods and it is straightforward to adjust the significance level for choosing the "most significant" of several independent tests. In general, if a is the smallest significance level of n independent tests, then the true overall significance level α is given by $\alpha = 1-(1-a)^n$. But, of course, there are 108 four-year periods if we are permitted to start in any month. Indeed, there is no reason to limit ourselves to four-year periods, since this choice was dictated by the data, and thus we should allow for tests calculated from one-year, two-year, and so on up to twelve-year periods.

Since test statistics based on partially overlapping subperiods will be correlated, there is no simple way of adjusting the nominal significance level to obtain the true significance level. It can be shown that an upper bound for the overall significance level of n tests at nominal level a is na. However, in this case, where n is very large and where the tests under consideration are obviously highly correlated, the use of this bound would severely overinflate the true significance level.

The general procedure adopted in this paper is to directly estimate the overall significance level using a Monte Carlo study. This technique is particularly suited to the problem at hand since, in the case of the sign test, there are no unknown parameters on which the test statistic depends. In order to carry out the technique, however, I must first choose an appropriate test statistic, and again this is not obvious. For any particular subperiod, the absolute value of the deviation of the number of positive signs from half the number of observations in the subperiod is the appropriate statistic. But, given this statistic for numerous subperiods of varying lengths from the overall sample, how do we combine them to obtain an overall test statistic? This also is a question best answered by a Monte Carlo experiment which leads to the two-step procedure which I now describe.

In step 1, 10,000 random samples of 155 plusses and minuses were constructed (corresponding to the number of monthly data points over the flexible rate period), where each plus or minus was generated with probability of one-half.[13] For each sample of 155 observations, the statistics $z_1, \dots, z_{12}$ were calculated where, for $k = 1, \dots, 12$ years,

$$z_k = \text{Max}\, N_k,$$

where N_k is the absolute value of the number of plusses less the expected number of $6k$ and where the maximization is carried out over every k year subperiod in that sample of 155 months. Thus z_k is the most extreme deviation from the null hypothesis found in any k year subperiod. The restriction to whole year subperiods was made to reduce the computational burden, but the subperiods were allowed to begin in any month. This step of the Monte Carlo study

[11] The pound floated against the dollar in June 1972. By May 1973 the dollar was floating against most other major currencies. It should also be noted that, since 1947, exchange controls in varying degrees have been applied to U.K. residents until controls were finally abolished in October 1979. However, non-U.K. residents have been free to switch between sterling and dollars since 1958 (A. R. Prest and D. J. Coppock, 1984, pp. 133, 152, and 169). Hansen and Hodrick (1983) argue that until January 1976 there existed a transitional period in which agents may have anticipated a return to fixed exchange rates. Any such arguments, which reduce the effective period of freely floating exchange rates, strengthen the test results found in this paper.

[12] The choice of monthly data was motivated by the need to find matching data for the exchange rate, forward rate, interest rates, and commodity prices.

[13] Random numbers were computed using the NAG library pseudo-random number generator. This has a cycle length of 2^{57}. A total of 3.1 million random numbers were required for this study.

TABLE 1—CUMULATIVE DISTRIBUTION OF z_k

	Subinterval Length: k Years											
z_k	1	2	3	4	5	6	7	8	9	10	11	12
0	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000
1	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000
2	10000	10000	10000	10000	10000	10000	9999	10000	9998	9987	9936	9663
3	9978	9996	9987	9992	9984	9980	9964	9927	9842	9667	9333	8561
4	8348	9694	9770	9744	9735	9656	9585	9376	9152	8729	8179	7273
5	3033	7789	8682	8818	8854	8731	8539	8314	7941	7418	6838	6043
6	353	4460	6384	7050	7257	7270	7203	6975	6570	6134	5581	4913
7	–	1808	3883	4957	5409	5568	5616	5420	5170	4875	4466	3957
8	–	566	1950	3037	3620	3967	4110	4063	3985	3802	3538	3100
9	–	119	801	1685	2290	2662	2863	2947	2916	2863	2673	2348
10	–	13	278	807	1307	1657	1902	2046	2081	2073	1977	1756
11	–	2	97	363	687	1000	1210	1368	1450	1429	1386	1268
12	–	0	28	132	334	584	748	868	961	973	971	891
13	–	–	11	42	139	292	423	520	604	648	653	588
14	–	–	3	20	58	145	218	294	348	392	429	400
15	–	–	0	4	23	60	111	158	199	243	268	260
16	–	–	0	1	7	22	51	80	112	144	158	163
17	–	–	0	0	2	7	23	37	62	84	93	98
18	–	–	0	0	0	3	8	19	27	48	58	50
19	–	–	–	0	0	0	4	13	15	25	37	30
20	–	–	–	0	0	0	3	4	6	12	15	21
21	–	–	–	0	0	0	1	3	3	6	10	8
22	–	–	–	0	0	0	0	1	3	3	4	6
23	–	–	–	0	0	0	0	1	1	1	2	4
24	–	–	–	0	0	0	0	0	1	1	1	3
25	–	–	–	–	0	0	0	0	0	0	1	1
26	–	–	–	–	0	0	0	0	0	0	0	1
27	–	–	–	–	0	0	0	0	0	0	0	0

Note: For each subperiod of k years, the table provides $a_k(z_k)$, the number of times the value z_k was attained or exceeded in 10,000 simulations of 155 months each.

thus generated 10,000 values of z_k for each $k=1,\dots,12$, and the cumulative distribution a_k, where for each k, $a_k(z_k)$ is the number of times z_k was obtained or exceeded, is provided in Table 1. Thus $a_k(z_k)$ represents the nominal significance level of test statistics based on subperiods of length k years.

Finally, for each sample, let

$$Y = \operatorname*{Min}_k a_k(z_k),$$

that is, where the minimization is carried out over the nominal significance levels of the z_k statistics for $k=1,\dots,12$. The statistic Y is the natural overall test statistic to use to look for alternatives H_1. The statistic Y represents the most extreme value for the 576 subperiods considered, with low values of Y indicating greater deviations from the null hypothesis. To calculate the true significance level of a test based on Y we need the distribution of Y. This is found in step 2. Thus 10,000 new random samples of 155 plusses and minuses were generated. Using Table 1, for each random sample the value of Y was calculated. The cumulative distribution of Y was then calculated and this is given in Table 2.

To carry out this test using data on x_t, $t=1,\dots,155$, one can calculate the z_k, $k=1,\dots,12$, from the signs of x_t and compute Y using Table 1. Table 2 then provides an estimate of the true significance level for the observed Y. Alternatively, one calculates z_k for a suspect subperiod and uses the tables to find the corresponding Y and an upper bound on the true significance level. This test is so constructed that full protection is provided against data mining within the sample. Furthermore, there is no small sam-

TABLE 2—SIGNIFICANCE LEVELS FOR DISTRIBUTION OF Y

y	Significance Level	y	Significance Level
0	0.0002	111	0.0269
1	0.0005	112	0.0283
2	0.0009	119	0.0335
3	0.0011	132	0.0369
4	0.0015	139	0.0396
6	0.0022	144	0.0412
7	0.0027	145	0.0427
8	0.0029	158	0.0452
10	0.0030	163	0.0480
11	0.0037	199	0.0513
12	0.0041	218	0.0534
13	0.0057	243	0.0553
15	0.0061	260	0.0594
19	0.0063	268	0.0620
20	0.0067	278	0.0707
21	0.0073	292	0.0743
22	0.0078	294	0.0776
23	0.0090	334	0.0814
25	0.0094	348	0.0854
27	0.0099	353	0.1070
28	0.0108	363	0.1106
30	0.0110	392	0.1156
37	0.0119	400	0.1213
42	0.0135	423	0.1248
48	0.0143	429	0.1267
50	0.0149	520	0.1312
51	0.0159	566	0.1486
58	0.0182	584	0.1538
60	0.0187	588	0.1602
62	0.0195	604	0.1643
80	0.0205	648	0.1675
84	0.0213	653	0.1708
93	0.0224	687	0.1773
97	0.0247	748	0.1832
98	0.0259	801	0.2012

Note: Significance level is the estimated probability of obtaining a value of Y less than or equal to y.

ple bias since the test is based on a Monte Carlo study rather than asymptotic results.[14]

B. *Discussion of Risk Adjustment*

Before describing the results of the test, I return to the issue of risk. Although, as mentioned in the introduction, there is evidence of a nonzero risk premium, I initially give the results of applying the test just described to the unadjusted series, $x_{t+1} = s_{t+1} - f_t$, for the following reasons:

(*i*) There is by no means a unanimity of opinion on the existence of risk premia. For example, Jeffrey Frankel (1982) estimated a model using a portfolio balance approach and was unable to reject the hypothesis of risk neutrality.

(*ii*) Tests rejecting simple efficiency are usually based on asymptotic results. Although large data sets are in fact being used, these typically involve several correlated exchange rates and overlapping data subject to serial correlation, making the effective sample size much smaller. If the underlying distribution has high kurtosis and, as suggested in this paper, is nonsymmetric, one would expect the small sample bias to be exacerbated. Related points are discussed by W. S. Krasker (1980), Craig Hakkio (1983) and Robert Korajczyk (1985).[15]

(*iii*) The deviations between the forward and expected future spot rates, usually attributed to risk premia, are small compared to the difference between forward and actual future spot rates. For example, Table 1 of Lars Hansen and Robert Hodrick (1983) obtained an $R^2 = 0.095$ in the case of the sterling-dollar exchange rate for a regression of $x_{t+1} = s_{t+1} - f_t$ on eight regressors. Furthermore there appears to be general agreement that the unconditional expected value of x_t is 0.

(*iv*) There is no agreement on the appropriate empirical model of the risk premium. Thus any specific adjustment for risk will be subject to some criticism.

(*v*) There does not appear to be any persuasive reason, under the null hypothesis of no bubbles, for believing that the pound sterling was considered to be less risky on average than the dollar over the 1981–84

[14] This may be a significant advantage. The small sample bias of asymptotic approximations to some nonparametric statistics can be important, particularly in the tails, and the continuity correction does not necessarily improve the approximation (see E. L. Lehman, 1975, pp. 16–17).

[15] Another technical problem concerns the treatment of nonstationarity in some tests. Considering the related issue of risk premia in the term structure of interest rates, Gary Shea (1985) found that the simple expectations hypothesis could not be rejected when the model was estimated using fractional difference methods.

subperiod. This is the direction of correction required if risk is to explain the predominance of negative returns to holding sterling during this period.

These considerations suggest that it would be best to apply my test both to the unadjusted series for excess returns and to a risk-adjusted series. In the remainder of this section, I describe the results of the test for the unadjusted data.

C. *Test Results*

Table 3 presents point in time monthly data, 1981–84, for the U.S. dollar price of the U.K. pound sterling and for the realized value of x_t, the excess return to holding a one-month forward contract to buy sterling, expressed in percent per month. The data were collected from the *Financial Times* and refer to the last working day of the month. See the Data Appendix for further details. For comparison, the x_t series was also computed from (2) using one-month Eurocurrency rates. The results (not shown) exhibited a close correspondence and, in particular, the sign of x_t for every month was identical for the two methods of calculation.

During this 48-month period, the x_t series was positive in 9 months and negative in 39 months, providing a value of $z_4 = 15$. From Table 1, we see that this corresponds to a value for the test statistic of $Y = 4$. The significance level of this result, obtained from Table 2, is estimated to be 0.0015. Thus the null hypothesis of a zero median for x_t over the full period of floating can be rejected, against the alternative of a negative median over the 1981–84 period, at very low significance levels.

II. Risk Premia

Can the results of the previous section be explained by risk premia?[16] The modeling of risk premia in foreign exchange has been investigated, inter alia, by Frankel (1982), Hansen and Hodrick (1983), Hodrick and Sanjay Srivastava (1984), Korajczyk (1985), and Nelson Mark (1985). For the purposes of this paper, the most convenient method for correcting for risk is given by Korajczyk. All of the other methods require the estimation of parameters using specific models of risk.[17]

Two-country dynamic general equilibrium models with risk-averse agents of the type examined by Robert Lucas (1982) and others generate expressions specifying how bond risk premia are incorporated into real interest rates and can be used to obtain formulae for the risk premium in the foreign exchange market. Provided deviations from purchasing power parity satisfy the martingale property, it can be shown that the risk premium separating the forward price of foreign currency and the expected future spot price is equal to the difference between the expected real interest rates on default-free nominal bonds denominated in the two currencies.[18]

The simple efficiency condition (1) is thus replaced by the condition

$$(4) \qquad E_t(s_{t+1} - f_t) = E_t r^*_{t+1} - E_t r_{t+1},$$

where $r_{t+1} = i_{t,1} - \Delta p_{t+1}$, $r^*_{t+1} = i^*_{t,1} - \Delta p^*_{t+1}$, and p_t and p^*_t are logarithms of domestic and foreign price indices. Korajczyk found

[16] Another conceivable explanation of the results is transactions costs. Given the size of the exchange rate movements over this period, the existence of transactions costs does not appear to be a credible explanation. The average excess return to holding dollar rather than sterling interest-bearing accounts over 1981–84 was 1.595 percent per month, i.e., an annual rate of 19.14 percent. Since brokerage costs for currency conversion are quite small for large sums of money, and since they need only have been incurred twice, transactions costs would clearly have been quite minor compared to the gain of such a switch for, say, a year or more.

[17] Furthermore, Frankel's estimates were consistent with risk neutrality; two models considered by Hansen and Hodrick incorporated restrictions rejected by the data; the third model considered by Hansen and Hodrick was found by Hodrick and Srivastava to be rejected by the data; and Mark's estimates of the key risk-aversion parameter show enormous variation in magnitude depending on the detailed implementation.

[18] This result, which is derived in Korajczyk, was also stated in Hodrick and Srivastava, fn. 5. Further references are provided in both papers.

TABLE 3—EXCESS RETURNS TO HOLDING STERLING

Date	$Spot_t$	x_t	$\tilde{x}_t$
12/80	2.3910		
01/81	2.3670	−1.40537	−1.25349
02/81	2.2050	−7.38488	−7.24246
03/81	2.2445	1.51735	2.52236
04/81	2.1405	−4.95351	−2.57442
05/81	2.0700	−3.79193	−3.51709
06/81	1.9305	−7.48292	−7.26054
07/81	1.8415	−5.28804	−5.47266
08/81	1.8490	−0.11892	0.33561
09/81	1.8050	−2.77553	−2.85732
10/81	1.8600	2.93514	3.63439
11/81	1.9550	5.06205	5.72062
12/81	1.9100	−2.00593	−1.96500
01/82	1.8810	−1.37277	−1.33888
02/82	1.8215	−3.22496	−3.53168
03/82	1.7820	−2.26375	−1.24831
04/82	1.7940	0.47493	2.20369
05/82	1.7905	−0.35124	−0.49467
06/82	1.7435	−2.77167	−3.63136
07/82	1.7380	−0.54225	−0.86245
08/82	1.7170	−1.26167	−1.40198
09/82	1.6945	−1.31618	−1.55394
10/82	1.6770	−1.08237	−0.83443
11/82	1.6235	−3.24520	−2.60923
12/82	1.6175	−0.29940	−0.16254
01/83	1.5200	−6.12127	−6.35737
02/83	1.5150	−0.16488	0.00571
03/83	1.4835	−1.85330	−1.96748
04/83	1.5605	5.14451	5.81921
05/83	1.6045	2.90563	2.66080
06/83	1.5340	−4.40917	−4.58661
07/83	1.5210	−0.85433	−0.73273
08/83	1.4940	−1.83383	−1.70438
09/83	1.4970	0.17049	0.14369
10/83	1.4955	−0.09023	−0.02817
11/83	1.4630	−2.23391	−2.02398
12/83	1.4515	−0.86091	−0.67521
01/84	1.4015	−3.56399	−4.12883
02/84	1.4905	6.11761	6.09640
03/84	1.4425	−3.34381	−3.19791
04/84	1.3985	−3.24669	−2.26886
05/84	1.3855	−1.13037	−0.87848
06/84	1.3565	−2.30640	−2.22165
07/84	1.3075	−3.93311	−4.12599
08/84	1.3085	0.12619	0.58454
09/84	1.2350	−5.86507	−6.09918
10/84	1.2210	−1.19270	−0.79855
11/84	1.1980	−1.82793	−1.59381
12/84	1.1590	−3.24697	−3.46695

Notes: Date is last working day of the month; $Spot_t$ is U.S. dollar price of the U.K. pound sterling; x_t is excess return to holding sterling calculated according to (1), expressed as percent per month; $\tilde{x}_t$ is risk-adjusted excess return to holding sterling calculated according to (5), expressed as percent per month.

that deviations from (1) were consistent empirically with the reformulation (4).

To conduct my test for speculative bubbles, allowing for risk premia, I should subtract expected real interest rate differentials from the series for x_t. Unfortunately, expected real interest rates are not observable. While they could be estimated by a time-series model, a simpler procedure is to note that (4) implies

$$(5)\qquad E_t\tilde{x}_{t+1}=0,$$

where $\quad \tilde{x}_{t+1}=s_{t+1}-f_t+r_{t+1}-r^*_{t+1},$

and that it is straightforward to apply the test of the preceding section to the risk-adjusted excess returns $\tilde{x}_{t+1}$, calculated using *ex post* real interest rates. The cost of conducting the test using *ex post* differentials in real interest rates rather than *ex ante* differentials is a possible reduction in the power of the test since $\tilde{x}_{t+1}$ includes noise from unforecastable changes in goods prices.[19] However, in view of the strong results of the preceding section, this potential cost would appear to be acceptable.

The final column of Table 3 gives the monthly series for $\tilde{x}_t$ over 1981–84, calculated according to (5) using one-month dollar and sterling Eurocurrency rates for $i_{t,1}$ and $i^*_{t,1}$ and one-month changes in the U.S. Consumer Price Index and the U.K. Retail Price Index for p_t and p^*_t.[20] The risk adjustment changes the sign of excess returns from negative to positive in 2 of the 48 months, so that $z_4=13$ for the 1981–84 subperiod.[21] According to Table 1, this corresponds to $Y=42$. The overall significance level, ob-

[19]Nominal interest rates for the coming period are included in the information set.

[20]As noted by Korajczyk, the price indices provide a better measure of goods prices midmonth, so that there is some mismatch of timing in the calculation of $\tilde{x}_t$.

[21]Again, there are two ways to calculate $\tilde{x}_t$. Using covered interest arbitrage, it follows that $\tilde{x}_{t+1}=s_{t+1}-s_t+\Delta p^*_{t+1}-\Delta p_{t+1}$, so that $\tilde{x}_t$ is simply the change in the real exchange rate. For comparison, $\tilde{x}_t$ was recalculated using this formula and found to match closely the series given in Table 3. For every month the sign of $\tilde{x}_t$ was the same for both methods of calculation.

tained from Table 2, is 0.0135. Thus, although the risk adjustment does lead to a higher significance level than when the test is applied to the unadjusted series, there is still very considerable evidence against the null hypothesis of a zero median for risk-adjusted excess returns and in favor of a negative median over the 1981–84 period.

III. Possible Interpretations

The results of the previous two sections constitute strong evidence, according to my definition, of an empirical speculative bubble in the dollar-sterling exchange rate over 1981–84. Apart from type I error, there are three possible explanations for the finding: (*i*) bootstrap equilibria (rational bubbles); (*ii*) nonsymmetric innovations in fundamentals; and (*iii*) nonrational or disequilibrium expectations.

These points can be made clearly by examining a simple exchange rate model, adapted from a continuous time model of Flood and Garber (1982), and used here for expositional purposes only. The economy is described by the following three equations:

$$m_t - p_t = a + by_t - ci_{t,1}, \tag{6}$$

$$i_{t,1} = i^*_{t,1} + \hat{s}_{t+1} - s_t, \tag{7}$$

$$p_t = p^*_t + s_t, \tag{8}$$

where m is the logarithm of the domestic money supply, y is the logarithm of domestic output, and other variables have been previously defined. Equation (6) is the money market equilibrium equation, and b and c are positive parameters. (7) is the arbitrage equation, and $\hat{s}_{t+1}$ denotes the expected value for the spot exchange rate in period $t+1$ held by agents in period t. Under rational expectations $\hat{s}_{t+1} = E_t s_{t+1}$. For simplicity, I assume zero transactions costs and risk neutrality. (8) states that purchasing power parity holds in every period.

Solving (6)–(8) for s_t, we obtain

$$s_t = \theta \hat{s}_{t+1} + v_t, \tag{9}$$

where $0 < \theta = c(1+c)^{-1} < 1$ and $v_t = (1+c)^{-1}(m_t - p^*_t - a - by_t + ci^*_{t,1})$. Under rational expectations, the solution which I have referred to as the "market fundamentals" solution is

$$\bar{s}_t = \sum_{i=0}^{\infty} \theta^i E_t v_{t+i}, \tag{10}$$

providing this sum converges. Assume that v_t is exogenous and has a (stationary or nonstationary) ARMA representation in a symmetric white noise innovation ε_t, that is, $F(L)v_t = G(L)\varepsilon_t$ where L is the lag operator and where ε_t is a white noise disturbance with median 0. The innovation in $\bar{s}_t$ is given by

$$\bar{s}_{t+1} - E_t \bar{s}_{t+1} = \theta^{-1} \sum_{i=1}^{\infty} \theta^i (E_{t+1} v_{t+i} - E_t v_{t+i}).$$

Since the revision in the expectation of v_{t+i} must simply be proportional to ε_{t+1}, that is, $E_{t+1} v_{t+i} - E_t v_{t+i} = \alpha_i \varepsilon_{t+1}$, we have

$$\bar{s}_{t+1} - E_t \bar{s}_{t+1} = \left[\theta^{-1} \sum_{i=1}^{\infty} \theta^i \alpha_i\right] \varepsilon_{t+1}.$$

It follows that $\bar{s}_{t+1} - E_t \bar{s}_{t+1}$ has median 0 and that

$$\Pr[x_{t+1} \geq 0] = \Pr[s_{t+1} - E_t s_{t+1} \geq 0] = 1/2,$$

where Pr stands for the probability conditional on information available at time t. Hence x_{t+1} has zero conditional median providing expectations are rational, the exchange rate is determined by the fundamentals solution, and innovations in the fundamentals are symmetric.

A. *Bootstrap Equilibria*

Consider, now, solutions to (9) other than (10). Blanchard and Watson point out that such solutions, which they term "rational bubbles," may result in a tendency towards runs (a nonzero median) though this is not necessarily the case. The rational expectations solutions to (10) can always be written

in the form $s_t = \bar{s}_t + c_t$ where c_t satisfies $c_t = \theta E_t c_{t+1}$. An asymmetry in the innovation of the bootstrap or bubble term, c_t, can lead to an asymmetry in the innovation of s_t. As an example they discuss the solution given by

$$c_t = \begin{cases} (\pi\theta)^{-1} c_{t-1} + \mu_t & \text{with prob. } \pi \\ \mu_t & \text{with prob. } 1-\pi \end{cases}$$

where μ_t satisfies $E_{t-1}\mu_t = 0$. It follows from their analysis that if μ_t is symmetric, and if the innovations in market fundamentals are symmetric, then the conditional median of x_t will be nonzero for $1/2 < \pi < 1$. Essentially the small probability of a large crash is balancing the likelihood of a continuation of the bubble. It should be noted that μ_t may either be a function of the innovation in the fundamentals, or be a variable which is wholly independent of the fundamentals (in which case the solution is often called a "sunspot" equilibrium).

B. *Asymmetric Fundamentals*

A second possibility is that the fundamentals themselves may have nonsymmetric innovations. Suppose that $v_t = v_t^* + \varepsilon_t$, where ε_t is a symmetric white noise disturbance, and that $\Delta v_t^* = g\omega_t$, where v_0^* is given, where $\omega_0 = 1$ and where

$$\omega_t = \begin{cases} \omega_{t-1} & \text{with prob. } \pi \\ 0 & \text{with prob. } 1-\pi. \end{cases}$$

The $0-1$ variable ω_t may be thought of as modeling a permanent regime change. Since $E_t v_{t+i+1}^* = E_t v_{t+i}^* + gE_t\omega_{t+i+1}$, we have $E_t v_{t+i}^* = v_t^* + g\pi(1-\pi)^{-1}(1-\pi^i)$, from which it follows that the fundamentals solution is given by

$$\bar{s}_t = \begin{cases} (1-\theta)^{-1} v_t^* + g\pi\theta(1-\theta)^{-1} & \\ \quad \times (1-\theta\pi)^{-1} + \varepsilon_t & \text{if } \omega_t = 1, \\ (1-\theta)^{-1} v_t^* + \varepsilon_t & \text{if } \omega_t = 0. \end{cases}$$

Once the permanent regime change to $\omega = 0$ has occurred, the distribution of $\bar{s}_t$ is symmetric. However, it is straightforward to show that $\Pr[\bar{s}_{t+1} - E_t\bar{s}_{t+1} \geq 0] > 1/2$ if $1/2 < \pi < 1$ as long as $\omega_t = 1$. Thus if agents believe, for example, that there is a small probability of a regime change in any given period, such as a permanent shift in monetary policy, but a high probability that such a change will eventually occur, then this will induce a highly nonsymmetric conditional distribution to $\bar{s}_t$ and hence a nonzero conditional median to x_t. In this case it would be natural to say that there is a bubble in the fundamentals.

C. *Nonrational Expectations*

The final possibility is that expectations are not rational in the strict sense of being conditionally unbiased.[22] The specific possibilities run the range from transitional deviations as agents learn about the model as in Taylor (1975) to the more dramatic departures of the sort described in Charles Kindleberger (1978).

Suppose that the fundamentals follow a random walk with drift, that is, $\Delta v_t = d + \varepsilon_t$, so that

$$\bar{s}_t = (1-\theta)^{-2}\theta d + (1-\theta)^{-1} v_t.$$

Suppose also that at time $t = 0$, the drift d, which is not directly observable, falls to d'. If expectations initially remain at $\hat{s}_{t+1} = s_t + (1-\theta)^{-1}d$ then $s_t = (1-\theta)^{-2}\theta d + (1-\theta)^{-1}v_t$ where $\Delta v_t = d' + \varepsilon_t$ and it follows[23] that $\Pr[x_{t+1} \geq 0] < 1/2$. If forecasts are revised on the basis of recent data as their bias becomes recognized, then expectations may eventually converge to the new rational forecast function. However, until convergence has been achieved, we can expect x_t to have a negative conditional median and mean.

The existence of such nonrational expectations does not require imperfect knowledge

[22]Allowing for risk, I would restate this last possibility as $E_t\tilde{x}_{t+1} \neq 0$. The views of John Bilson (1981) seem to belong to this category. Bilson has argued that the foreign exchange markets are inefficient in the sense that trading strategies can be developed which are too profitable to be explained by compensation for risk.

[23]I assume that $f_t = \hat{s}_{t+1}$ so that $x_{t+1} = s_{t+1} - \hat{s}_{t+1}$.

of the true model. Suppose that, at $t=0$, d is reduced to $d=0$ as a result of a government policy which is announced, fully credible, and actually adhered to. But, suppose that individually rational agents are not confident that other agents will immediately and fully adjust their expectations, either because they do not believe that other agents are rational (or that other agents believe that *they* are rational), or because they are not confident that other agents find the announced policy credible. Then again expectations will be biased, and, although they may well converge to equilibrium, it is likely that the mean and median of x_t will be negative during the transition. Arguments of this sort are extensively discussed in the volume edited by Roman Frydman and Edmund Phelps (1983).

As a final example, suppose that v_t is symmetric white noise and that agents believe the spot rate is following the path

$$s_t = \begin{cases} \beta + \delta t + v_t & \text{for } t = 0,\dots,T-1 \\ v_t & \text{for } t \geq T. \end{cases}$$

With $\hat{s}_{t+1}$ computed on this assumption, the actual path followed by s_t will be

$$s_t = \begin{cases} \theta(\beta+\delta) + \theta\delta t + v_t & \text{for } t = 0,\dots,T-2 \\ v_t & \text{for } t \geq T-1. \end{cases}$$

It is easy to check that the conditional median and mean of x_t will generally deviate from zero for $t = 0,\dots,T-1$. However, for θ near 1, δ small and appropriate choice of β, the expectations of agents will not be badly erroneous except that the price will collapse one period earlier than expected. This example suggests that there may be a wide class of "near rational" solutions.[24]

D. *Discussion*

Is it possible to distinguish between the three types of bubbles just described? It would appear to be very difficult. One might hope to detect nonrational bubbles by tests based on the mean of x_t. However, the detection of a nonzero mean will be difficult for two reasons. First, if the deviation from rationality occurs over only part of the sample period, then techniques analogous to those of this paper must be developed for the sample mean. Second, most tests based on the mean assume a symmetric distribution, yet rationality is consistent with skew distributions. Furthermore, it is clear that a highly skew distribution of x_t, whether the result of a bootstrap solution or of a bubble in the fundamentals, can statistically mimic irrationality. Since an apparent bias over a limited period of time can always be explained by the possibility of a rare large event of opposite sign, it may be impossible to develop a statistically powerful test to detect transitory irrationality.[25] Indeed, for conjectured rare or unique events, it may be impossible to define the objective probability distribution, making it difficult to distinguish conceptually between rational and irrational expectations.

Distinguishing between the types of rational bubbles also presents considerable difficulties. In a somewhat different context, it has been argued by James Hamilton and Charles Whiteman (1985) that apparent evidence of what I have referred to as a bootstrap equilibrium can be reinterpreted as having arisen from a suitable specification of fundamentals which are unobserved by the econometrician. The above examples illustrate this point for the case at hand. The negative median for x_t over 1981–84 could equally well be explained by a bootstrap bubble in which the run of negative returns on holding sterling was compensated by the possibility of a collapse of the dollar or, say, by the discounting of a strong U.K. output boom, expected to eventually occur but not fully realized during the period. Distinguish-

[24]An interesting question is whether there exist near rational solutions in the more precise sense of George Akerlof and Janet Yellen (1985).

[25]There appears to be little point examining the x_t data for skewness over 1981–84. A finding of skewness would still be compatible with irrationality. An apparently symmetrical empirical distribution would still be compatible with theoretical skewness if the latter is due to rare possibilities not actually observed.

ing empirically could be particularly difficult since a bootstrap equilibrium can be a function solely of fundamentals.

Nevertheless, imposing rationality is a restriction and assuming a fundamentals solution is an additional restriction. What is clear from the above discussion is that distinguishing between the competing explanations for a bubble given above can only be attempted within the context of a complete model of the exchange rate.[26]

IV. Conclusions

This paper has developed a test for a type of abnormality in excess returns which we have taken to define a bubble. In particular, I have defined a speculative bubble as occurring during any subperiod in which the (risk-adjusted) excess return on holding an asset has a nonzero median. This definition, while not conventional, is a natural use of the term, since it implies a tendency toward an abnormal number of positive or negative excess returns and has the advantage of being directly testable from data on excess returns. Indeed, with speculative bubbles so defined in terms of this measure of the central tendency of excess returns, it is possible to test for bubbles using nonparametric methods and without assuming any particular model of the exchange rate. The principal statistical complication, treated in this paper, is the necessity to develop a procedure for detecting a nonzero median over only part of the whole sample period. Equivalently, we may think of this as a procedure which controls for data mining.

When applied to the sterling-dollar exchange rate this test detects the presence of a negative median excess return on holding sterling rather than dollar assets over the 1981–84 period. After carefully controlling for possible data mining from the full period of flexible rates, the significance level of the test is 0.0015 when using unadjusted excess returns and 0.0135 when an allowance is made for possible risk premia. The finding of a bubble during 1981–84 can be explained by either nonrational expectations, bootstrap equilibria (i.e., rational bubbles) or nonsymmetric fundamentals. Distinguishing between these alternatives would be difficult and will require a complete model of the exchange rate (which will then become part of the maintained hypothesis). An intriguing open question is whether some variable can be found, whether interpreted as a fundamental or as a sunspot variable, which can explain the bubble found for the sterling-dollar rate during this period.

Many other questions remain to be investigated. The sterling-dollar rate was chosen for analysis because of the considerable discussion among policymakers in each country, at various times, suggesting that they believed the exchange rate between the currencies was out of line with its equilibrium value. Will empirical speculative bubbles be detected by these methods between other currencies and in other assets?[27] It would be desirable to extend this type of test to a multivariate context so that the exchange rate system could be analyzed as a unit. More powerful versions of the test could be devised which make use of the magnitudes of the changes or consider alternatives of multiple bubbles. Such extensions are nontrivial and must await further research.

Data Appendix

Except as noted below, the spot U.S. dollar price of the U.K. pound sterling, its

[26]After completing this paper, I became aware of Richard Meese (forthcoming) and Frankel and Kenneth Froot (1985). Meese considers various evidence for rational bubbles (i.e., bootstrap equilibria) in foreign exchange markets, focusing on a specification test which is conditioned on a particular model of the exchange rate. Frankel and Froot construct an account of the 1981–84 period, partly in terms of irrational speculation, which provides one possible explanation of the empirical findings of this paper.

[27]It should be noted that if bubbles are of short duration, they will be difficult to detect. For example, if bubbles last 12 months, and only appear once every 13 years, then they will be virtually undetectable using monthly data. From Tables 1 and 2, it can be seen that the significance level of a run of twelve negatives is only 0.107. The data spoke clearly in this case because the bubble lasted 4 years.

one-month forward premium, and the sterling and U.S. dollar one-month Eurocurrency interest rates were taken from issues of the *Financial Times* of London and refer to the last working day of the month. Exchange rate data were taken from the table labelled "Pound Spot and Forward" and refer to the market-closing price. Interest rate data were taken from the "Euro-Currency Interest Rates" table and refer to market-closing rates. In each case the figures used were the averages of the ranges given. The *Financial Times* was not published from June 1 to August 8, 1983, due to a printing dispute. For May 31, June 30, and July 29, 1983, equivalent figures were taken from the London *Times*. The U.S. Consumer Price Index for all urban consumers was taken from the April 1985 issue of *Business Conditions Digest*, Table 320, and the U.K. Index of Retail Prices for All Items was taken from Central Statistical Office publication *Economic Trends*, issues for May 1982 (Table 18.8), May 1983 (Table 18.8), and May 1985 (Table 18.1).

REFERENCES

Akerlof, George A. and Yellen, Janet L., "Can Small Deviations from Rationality Make Significant Differences to Economic Equilibria?," *American Economic Review*, September 1985, *75*, 708–20.

Baillie, Richard T., Lippens, Robert E. and McMahon, Patrick, "Testing Rational Expectations and Efficiency in the Foreign Exchange Market," *Econometrica*, May 1983, *51*, 553–63.

Bilson, John F. O., "The 'Speculative Efficiency' Hypothesis," *Journal of Business*, July 1981, *54*, 435–52.

Blanchard, Olivier J., "Speculative Bubbles, Crashes and Rational Expectations," *Economic Letters*, 1981, *4*, 387–89.

______ **and Watson, Mark,** "Bubbles, Rational Expectations and Financial Markets," in P. Wachtel, ed., *Crises in the Economic and Financial Structure*, Lexington: Lexington Books, 1982.

Burmeister, Edwin, Flood, Robert P. and Garber, Peter M., "On the Equivalence of Solutions in Rational Expectations Models," *Journal of Economic Dynamics and Control*, May 1983, *5*, 311–21.

Diba, Behzad and Grossman, Herschel I., "Bubbles, Rational Expectations and Financial Markets," NBER Working Paper No. 1059, 1983.

______ **and** ______, "Rational Bubbles in the Price of Gold," NBER Working Paper No. 1300, 1984.

Dixon, W. J. and Mood, A. M., "The Statistical Sign Test," *Journal of the American Statistical Association*, 1946, *41*, 557–66.

Dornbusch, Rudiger, "Expectations and Exchange Rate Dynamics," *Journal of Political Economy*, December 1976, *84*, 1161–76.

Edwards, Sebastian, "Floating Exchange Rates, Expectations and New Information," *Journal of Monetary Economics*, May 1983, *11*, 321–36.

Evans, George W., "Expectational Stability and the Multiple Equilibria Problem in Linear Rational Expectations Models," *Quarterly Journal of Economics*, November 1985, *100*, 1217–33.

Flood, Robert P. and Garber, Peter M., "Market Fundamental versus Price Level Bubbles: The First Tests," *Journal of Political Economy*, August 1980, *88*, 745–70.

______ **and** ______, "Bubbles, Runs, and Gold Monetization," in P. Wachtel, ed., *Crises in the Economic and Financial Structure*, Lexington: Lexington Books, 1982.

Frankel, Jeffrey A., "In Search of the Exchange Risk Premium: A Six-Currency Test Assuming Mean Variance Optimization," *Journal of International Money and Finance*, December 1982, *1*, 255–74.

______ **and Froot, Kenneth,** "The Dollar as an Irrational Speculative Bubble: A Tale of Fundamentalists and Chartists," mimeo., University of California-Berkeley, December 1985.

Frenkel, Jacob A., "The Forward Exchange Rate, Expectations, and the Demand for Money: The German Hyperinflation," *American Economic Review*, September 1977, *67*, 653–70.

______, "Flexible Exchange Rates, Prices, and the Role of 'News': Lessons from the 1970s," *Journal of Political Economy*, August 1981, *89*, 655–705.

______, *Exchange Rates and International Macroeconomics*, NBER, Chicago: University of Chicago Press, 1983.

Frydman, Roman and Phelps, Edmund S., *Individual Forecasting and Aggregate Outcomes: 'Rational Expectations' Examined*, New York: Cambridge University Press, 1983.

Geweke, John and Feige, Edward, "Some Joint Tests of the Efficiency of Markets for Forward Foreign Exchange," *Review of Economics and Statistics*, August 1979, *61*, 334–41.

Gourieroux, C., Laffont, J. J. and Monfort, Alain, "Rational Expectations in Linear Models: Analysis of Solutions," *Econometrica*, March 1982, *50*, 409–25.

Hakkio, Craig S., "Expectations and the Forward Exchange Rate," *International Economic Review*, October 1981, *22*, 663–78.

______, "Comment" on L. P. Hansen, and R. J. Hodrick, in J. A. Frenkel, ed., *Exchange Rates and International Macroeconomics*, Chicago: University of Chicago Press, 1983.

Hamilton, James D. and Whiteman, Charles H., "The Observable Implications of Self-fulfilling Expectations," *Journal of Monetary Economics*, November 1985, *16*, 353–74.

Hansen, Lars Peter and Hodrick, Robert J., "Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis," *Journal of Political Economy*, October 1980, *88*, 829–53.

______ **and** ______, "Risk Averse Speculation in the Forward Foreign Exchange Market: An Economic Analysis of Linear Models," in J. A. Frenkel, ed., *Exchange Rates and International Macroeconomics*, Chicago: University of Chicago Press, 1983.

Hodrick, Robert J. and Srivastava, Sanjay, "An Investigation of Risk and Return in Forward Foreign Exchange," *Journal of International Money and Finance*, April 1984, *3*, 5–29.

Kindleberger, Charles, P., *Manias, Panics and Crashes: A History of Financial Crises*, New York: Basic Books, 1978.

Korajczyk, Robert A., "The Pricing of Forward Contracts for Foreign Exchange," *Journal of Political Economy*, April 1985, *93*, 346–68.

Krasker, W. S., "The 'Peso Problem' in Testing the Efficiency of Forward Exchange Markets," *Journal of Monetary Economics*, April 1980, *6*, 269–76.

Lehman, E. L., *Nonparametrics*, San Francisco: Holden-Day, 1975.

LeRoy, Stephen F., "Efficiency and the Variability of Asset Prices, *American Economic Review Proceedings*, May 1984, *74*, 183–87.

Lucas, Robert E., Jr., "Interest Rates and Currency Prices in a Two-Country World," *Journal of Monetary Economics*, November 1982, *10*, 335–60.

McCallum, Bennett T., "On Non-uniqueness in Rational Expectations Models: An Attempt at Perspective," *Journal of Monetary Economics*, March 1983, *11*, 139–68.

McKinnon, Ronald, I., *Money in International Exchange: The Convertible Currency System*, New York: Oxford University Press, 1979.

______, *An International Standard for Monetary Stabilization*, Washington: Institute for International Economics, 1984.

Mark, Nelson C., "On Time Varying Risk Premia in the Foreign Exchange Market; An Economic Analysis," *Journal of Monetary Economics*, July 1985, *16*, 3–18.

Meese, Richard A., "Testing for Bubbles in Exchange Markets: A Case of Sparkling Rates?," *Journal of Political Economy*, forthcoming.

Prest, A. R. And Coppock, D. J., *The UK Economy*, 10th ed., London: Weidenfeld and Nicolson, 1984.

Shea, Gary S., "Long Memory Models of Interest Rates, the Term Structure, and Variance Bounds Tests," International Finance Discussion Paper No. 258, Federal Reserve Board, July 1985.

Shiller, Robert J., "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?," *American Economic Review*, June 1981, *71*, 421–36.

Taylor, John B., "Monetary Policy during a Transition to Rational Expectations," *Journal of Political Economy*, October 1975, *83*, 1009–22.

______, "Conditions for Unique Solutions in Stochastic Macroeconomic Models with

Rational Expectations," *Econometrica*, September 1977, *45*, 1377–85.

West, Kenneth D., (1985a) "Dividend Innovations and Stock Price Volatility," mimeo., Princeton University, 1985.

———, (1985b) "A Specification Test for Speculative Bubbles," mimeo., Princeton University, 1985.

United Kingdom, Central Statistical Office, *Economic Trends*, London, England.

U.S. Bureau of Economic Analysis, *Business Conditions Digest*, Washington, D.C.

The Rigidity of Prices

By DENNIS W. CARLTON*

For many transactions, prices remain rigid for periods exceeding one year. Price rigidity is positively correlated with industry concentration. For several products, the correlation of price changes across buyers is low. The paper also investigates the relationship between price rigidity, price change, and the length of time a buyer and seller have been doing business. The evidence emphasizes the importance of nonprice rationing and the inadequacy of models in which price movements alone clear markets.

Economists focus on price as a mechanism to allocate resources efficiently. It is well recognized that inefficient resource allocation could occur if prices are not free to adjust. Much of macroeconomics relies on some, usually unexplained, source of price rigidity to generate inefficient unemployment. And in industrial organization there is a large literature on "administered" prices which fail to respond to the forces of supply and demand. Recently, there have been several attempts to explain price rigidity (see, for example, Arthur Okun (1981) and Oliver Williamson (1975)) and to develop a theory to explain why efficient resource allocation requires price to be unchanging or "rigid" (see, for example, my forthcoming paper and Robert Hall, 1984). Whether or not price rigidity is efficient, one common conclusion emerging from models with price rigidity is that markets with rigid prices behave very differently than markets with flexible prices. Therefore, an important unanswered question is, just how rigid are prices? Despite the great interest in this question, there have been virtually no attempts to answer it with data on individual transaction prices.

The purpose of this paper is to present evidence on the amount of price rigidity that exists in individual transaction prices. Previous studies of price rigidity have relied almost exclusively on an examination of aggregate price indices collected by the Bureau of Labor Statistics (BLS).[1] The use of BLS data has been strongly criticized on the grounds that the BLS data are inaccurate measures of transaction prices. George Stigler and James Kindahl sought to remedy this deficiency by collecting price data on actual transactions. Stigler and Kindahl then showed that price indices of average transaction prices were more flexible than the BLS price indices.

The difficulty with using indices is that they can mask the behavior of individual transaction prices. For example, suppose that two people buy varying amounts of commodity *A* monthly for many years. Suppose that each buyer pays a constant price on each transaction for a period of several years, that when the price to one buyer changes, the price to the other buyer is unaffected and that the price rigidity that exists is more pronounced for a downward price movement. All of these facts could be perfectly consistent with a flexible aggregate price in-

*Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, IL 60637 and National Bureau of Economic Research. I thank the NSF and the Law and Economics Program at the University of Chicago for support. I thank Frederic Miller, Virginia France, Larry Harris, Deborah Lucas, and Steven Oi for research assistance. I also thank Claire Friedland and George Stigler for making these data available to me and for assisting me in their use. I thank Edward Lazear, Sam Peltzman, George Stigler, two anonymous referees, and participants at seminars at the NBER, Stanford, the universities of Chicago, Montreal, Pennsylvania, and Virginia for helpful comments.

[1]Research on prices includes the early and important work of Frederick Mills (1926), Gardiner Means (1935), and more recently, George Stigler and James Kindahl (1970).

dex as long as the amount purchased by each buyer varies from month to month. Yet the implication that many draw from a flexible price index, namely that price is allocating resources efficiently, could be completely inappropriate. Moreover, there are several interesting questions that cannot be answered by examining aggregate price indices. For example, how long do prices to a buyer remain unchanged, what is the relationship between contract length and price rigidity, and how closely together do the prices to different buyers move?

Using the Stigler-Kindahl data, I have examined the behavior of individual buyers' prices for certain products used in manufacturing. My main conclusions are:

1) The degree of price rigidity in many industries is significant. It is not unusual in some industries for prices to individual buyers to remain unchanged for several years.

2) Even for what appear to be homogeneous commodities, the correlation of price changes across buyers is very low.

3) There is a (weak) negative correlation between price rigidity and length of buyer-seller association. The more rigid are prices, the shorter the length of association.

4) There is a positive correlation between price rigidity and average absolute price change. The more rigid are prices, the greater is the price change when prices do change.

5) There is a negative correlation between length of buyer-seller association and average absolute price change. The longer a buyer and seller deal with each other, the smaller is the average price change when prices do change.

6) There is no evidence that there is an asymmetry in price rigidity. In particular, prices are not rigid downward.

7) The fixed costs of changing price at least to some buyers may be small. There are plenty of instances where small price changes occur. It appears that, for any particular product, the fixed cost of changing price varies across firms and buyers.

8) There is at best very weak evidence that buyers have systematic preferences across products for unchanging prices.

9) The level of industry concentration is strongly correlated with rigid prices. The more concentrated the industry, the longer is the average spell of price rigidity.

The most startling finding to me is that for many products, the correlation of price changes across buyers is low. Some of the theories referred to earlier explain why this is likely to occur, especially for specialized goods. The fact that it occurs for what most economists (though not necessarily businessmen) would regard as a homogeneous product emphasizes how erroneous it is to focus attention on price as the exclusive mechanism to allocate resources. Nonprice rationing is not a fiction, it is a reality of business and may be the efficient response to economic uncertainty and the cost of using the price system. (See my forthcoming paper.)

Two general caveats deserve mention. First, a rigid price, by itself, does not necessarily imply an inefficiency. If supply and demand are unchanging, prices will be rigid. Moreover, even in a changing market, a fixed-price contract for a fixed quantity creates no economic inefficiency in the standard competitive model. If prices change subsequent to the signing of the contract, the buyer incurs a capital gain or loss, but his marginal price remains the same as every other buyer as long as the product can be readily bought and sold. However, if either the buyer cannot readily resell his product, or if the buyer does not have a fixed quantity contract, then a fixed price may well lead to buyers facing different marginal prices. My understanding of the data I use is that the contracts typically leave the quantity unspecified, so that different buyers paying different prices do indeed face different marginal prices. Although this is inefficient in the standard competitive model, it need not be under more realistic assumptions that recognize the cost of making a market. (See my forthcoming paper. See also my 1978, 1979, 1982, 1983 papers for analyses reconciling observed price behavior with market equilibrium.) But the finding of different prices and price movements to different buyers does emphasize the inadequacy of the simple market-clearing model.

Second, the time period I examine is one with relatively low levels of inflation and therefore I have made no adjustment for it. However, even if inflation were rampant and all prices indexed so that no (nominal) price rigidity existed, the main conclusion of the paper would stand as long as some of the other empirical findings (such as the low correlation of price movements across buyers) continue to hold. The conclusion is that price alone is not allocating goods and that new theories are required to justify what looks like non-market-clearing behavior.

This paper is organized as follows. Section I describes the Stigler-Kindahl data and discusses measures of price rigidity. Section II analyzes the characteristics of price rigidity found in several general product groupings. Section III investigates the relationship between price rigidity, price change, and length of buyer-seller associations. Section IV examines whether buyers have systematic preferences for price stability across different products. One criticism of using broadly defined product groups as the unit of analysis is that there is so much heterogeneity of products within a single product grouping that results can be biased. Therefore, in Section V, I redo the analysis for a select group of narrowly defined products. Section VI shows how to measure whether the prices to different buyers move in concert and classifies the various products according to how similar are price changes to different buyers. Section VII examines some specific implications the results have for the prediction of price behavior. Section VIII examines whether there is any relationship of the various characteristics of price movements to the industry's structural characteristics.

I. The Stigler-Kindahl Data

Stigler and Kindahl collected data mainly from buyers on actual transaction prices paid for a variety of products. They tried to correct for any explicit or implicit discounting and for any changes in the specifications of the product. Although there is undoubtedly some misreporting of prices, and some unrecorded product changes (for example, physical characteristics, point of delivery, time of delivery), it is the most accurate and comprehensive data I know of on individual transaction prices.

The buyers who report prices are typically firms in the *Fortune* 500. The identity of the seller is not known.[2] Typically, there is only scant information on quantity purchased, though it is believed that during the course of the reporting buyers were using the product regularly. Ideally, actual transaction prices are reported monthly. However, in several instances, prices are reported less frequently. A decision on how (or whether) to interpolate prices had to be made.

If the price is unchanged between reportings, I assume that the intervening price is also unchanged. If the price is not the same, then I create two different series. One method assumes a change in each unobserved month. The other assumes only one change over the entire period. For example, suppose that for January, the price is \$10, and for April, it is \$20 with missing reports for February and March. The first interpolation approach assumes that the price was \$13.33 in February and \$16.67 in March (i.e., linear interpolation), while the second interpolation approach assumes that the price changed to \$20 in either February, March, or April. (It turns out that the results on length of rigidity are unaffected by which particular month is assumed for the price change in this second approach.)

The period of observation is January 1, 1957 through December 31, 1966. Few associations between buyers and seller last for

[2] The form in which the data exist do not allow conclusive determination that the buyer is dealing with only one seller. However, it is believed that only one seller is involved when the buyer is reporting prices pursuant to a contract. Furthermore, when prices remain unchanged or when the specification of the good remains unchanged from observation to observation, the buyer is also likely to be dealing with only one seller. I thank Claire Friedland, who helped collect the original data, for helpful discussions on this matter. For expositional ease, I will regard each price series as arising from a transaction between one buyer and one seller. I will point out when this assumption would substantially alter the interpretation of the results.

the entire ten-year period, a point which I analyze later on. Transactions often take place under "contract" and the length of the contract (for example, semiannual, annual) is indicated. The Appendix provides additional information on each type of transaction. Many contracts specify neither a price nor quantity. They seem not to be binding legal documents, but rather more like agreements to agree.

The commodities chosen for study were preselected by Stigler and Kindahl to contain many that others had claimed were characterized by inflexible prices. The commodities are intermediate products used in manufacturing. Within broad commodity classes, finer product distinctions are made. So, for example, one can examine the general category of steel or a specific product category like carbon steel pipe less than 3 inches in diameter. Even within fine product specifications, the individual transactions will probably not involve perfectly homogeneous goods. Therefore, I never compare absolute price levels across products, but instead look only at percentage changes in price and compare movements in percentage changes in price across buyers.

There are a few instances where price series are believed to be list prices, and those prices have been excluded from the analysis. Also excluded are price series that contain inconsistent information. For example, a series is excluded if the reporter claims to produce prices through 1965, but instead prices only through 1960 appear. For several transactions, the product undergoes a specification change. When this occurs, I treat the prices under the new specification change as a new transaction.

II. Analysis of Product Groups

Table 1 describes the price rigidity present in the individual transaction prices by product group. The first column in Table 1 lists the type of product purchased. Column 2 lists the number of buyer-seller pairings that are observed for goods of unchanged specification. (One pairing could last anywhere from 1 month to 10 years.) Column 3 lists the average duration of price rigidity. This last figure is computed as the average length of spell for which price remains unchanged. For example, if the observations on monthly price were \$5, \$5, \$5, \$6, \$6, \$7, \$7, \$7, \$7, the average rigidity would be three months. The procedure for calculating an average rigidity actually involves an underestimate since the price before the period of observation may have been \$5 and the price after the period of observation may have been \$7. Calculations including and excluding the beginning and ending spells were done with no material change in the substantive interpretation of the results. The calculations in Table 1 are based on the second method of interpolation of prices (only one price change between missing observations—see Section I) and include the beginning and the end of each price series. Column 4 reports the standard deviation in the rigidity of prices. Column 5 reports the same estimate of price rigidity as in column 3, except that only "monthly" contract series are used. These series have fewer missing observations than the other types of transactions, hence much less interpolation is needed. (In fact, the results on rigidity for monthly contracts are similar for the two methods of interpolation.) If the implication of the numbers in column 3 across commodities differ greatly from those in column 5, one might be suspicious of the interpolation used in column 3. I expect price flexibility of monthly contracts to exceed that of all other contract types, so column 5 really puts a lower bound on column 3.

To avoid misinterpretation of the results, it may be helpful to review a standard issue in duration analysis. Imagine that there are two observed transactions, each lasting for a one-year period and each involving the same size of monthly purchase. The first transaction involves a different price each month, while the second involves the same price each month. There are 13 spells of rigidity, 12 of which last one month and one of which lasts twelve months. Based on spells, the average rigidity is 24 ÷ 13 or 1.8 months with 92 percent of the spells lasting one month and 8 percent lasting twelve months. Conditional on a price change just having occurred, the average time to the next price

TABLE 1—PRICE RIGIDITY BY PRODUCT GROUP

Product Group (1)	Number of Buyer-Seller Pairings[a] (2)	Average Duration of Price Rigidity (Spells) (Months) (3)	Standard Deviation of Duration (Spells) (Months) (4)	Average Duration of Price Rigidity Monthly Contracts (Spells) (Months) (5)	Average Duration of Price Rigidity (Transactions) (Months) (6)
Steel	348	13.0	18.3	9.4	17.9
Nonferrous Metals	209	4.3	6.1	2.8	7.5
Petroleum	245	5.9	5.3	2.5	8.3
Rubber Tires	123	8.1	12.0	7.8	11.5
Paper	128	8.7	14.0	8.8	11.8
Chemicals	658	12.8	10.7	9.6	19.2
Cement	40	13.2	14.7	5.6	17.2
Glass	22	10.2	12.1	8.5	13.3
Truck Motors	59	5.4	6.3	3.7	8.3
Plywood	46	4.7	7.7	1.2	7.5
Household Appliances	14	3.6	3.6	2.5	5.9

[a]A "pairing" means a transaction over time for a good of constant specification.

change is 1.8 months. Yet, one-half of all goods sold involve a rigid price over the entire period. In other words, holding monthly purchases constant, the analysis based on spells underestimates the fraction of goods sold with rigid prices. The results in columns 3 and 5 utilize spells data. Therefore, even though I have no quantity information, I expect that these results underestimate the fraction of goods sold at rigid prices.

In column 6, I calculate price rigidity using a transaction as the unit of analysis, not a "spell." For each transaction, I calculate the average price rigidity, and then take an average (with each transaction weighted according to its length) over all transactions. Return to the earlier example of two transactions, each lasting one year, but one involving 12 price changes and the other no price changes. An analysis based on *transactions* (not spells) would calculate average rigidity to be $(1+12)/2$ or 6.5 months. It is that type of calculation that is reported in column 6.

Several interesting facts emerge from Table 1. In several industries, prices are on average unchanged over periods exceeding one year. The degree of price inflexibility varies enormously across products groups. Steel, chemicals, and cement have average rigidities exceeding one year while household appliances, plywood, and nonferrous metals have average price rigidities of less than five months. For any one product group the standard deviation of rigidity is quite high. In fact, the standard deviation tends to rise as the average duration of rigidity rises. The simple correlation and the Spearman Rank Correlation between the standard deviation and the average duration (cols. 3 and 4) are both above .80. This suggests (though does not prove) either that each product group presented in Table 1 contains heterogeneous products which differ widely in their price flexibility or that for even a homogeneous product a great heterogeneity in price flexibility is present.[3]

Column 5 shows that using monthly contracts rather than all contracts does not change the basic implications of column 3 regarding price rigidity across groups. Column 6 shows that, as expected, the average of price rigidity rises when the unit of analysis is a transaction. Indeed, the results of

[3]An alternative explanation is that price movements for the same product are similar across different transactions at any one instant but not across time. As we will see in Section VI, this explanation will turn out to be incorrect.

TABLE 2—FREQUENCY OF DURATION OF PRICE RIGIDITY FOR VARIOUS TYPES OF TRANSACTIONS BASED ON SPELLS OF PRICE RIGIDITY[a]

Product	Type of Transaction	Percent of all Transactions	Number of Pairings[b]	0–3 Months	4 Mo.–1 Year	1–2 Years	2–4 Years	Over 4 Years
Steel	Annual	3	11	.11	.41	.24	.22	.03
	Quarterly	53	185	.34	.26	.18	.12	.09
	Monthly	32	111	.48	.27	.15	.07	.04
Nonferrous Metals	Annual	4	8	.16	.69	.12	.03	0
	Quarterly	19	40	.61	.29	.08	.02	.02
	Monthly	42	87	.78	.20	.02	.01	0
Petroleum	Annual	27	66	.20	.69	.07	.04	0
	Quarterly	15	37	.74	.23	.02	.00	–
	Monthly	7	16	.83	.15	.02	0	–
Rubber Tires	Annual	26	32	.19	.72	.07	.01	.01
	Quarterly	37	45	.34	.48	.11	.04	.04
	Monthly	20	24	.44	.44	.07	.01	.06
Paper	Annual	17	22	.04	.69	.18	.08	.01
	Quarterly	2	3	.17	.42	.29	.08	.04
	Monthly	28	36	.46	.36	.12	.04	.02
Chemicals	Annual	43	286	.11	.58	.17	.09	.06
	Quarterly	11	72	.37	.30	.12	.16	.04
	Monthly	20	134	.53	.27	.09	.06	.04
Cement	Annual	20	8	.04	.78	.13	.04	0
	Quarterly	50	20	.19	.27	.23	.14	.05
	Monthly	10	4	.64	.29	.02	.04	.02
Glass	Annual	36	8	0	.87	.10	.03	0
	Quarterly	9	2	.25	.50	.19	0	.06
	Monthly	41	9	.51	.22	.18	.09	0
Truck Motors	Annual	14	8	.05	.86	.09	0	0
	Quarterly	2	1	.21	.57	.21	0	0
	Monthly	58	34	.69	.26	.04	.01	0
Plywood	Annual	0	0	0	0	0	0	0
	Quarterly	96	44	.64	.29	.04	.02	.01
	Monthly	4	2	.99	.02	0	0	0
Household Appliances	Annual	21	3	0	.82	.18	0	0
	Quarterly	0	0	0	0	0	0	0
	Monthly	57	8	.78	.22	0	0	0

[a] The numbers in the rows of the table may not add to one because of rounding.

[b] The "Number of Pairings" is not the number of spells of price rigidity in all contracts. See the discussion preceding Table 1, and the footnote to Table 1.

column 6 are striking in that they show that every product group has an average rigidity in excess of roughly six months, and that 6 of the 11 product groups have average rigidities of roughly one year or more.

In Table 2, more detailed evidence is provided on the time pattern of price rigidity by product group for three types of transactions. The three transaction types are monthly, in which case the transaction occurred monthly (with no necessary future commitment), quarterly monthly in which case the transaction was monthly but was observed quarterly, and annual in which case the transaction was pursuant to an annual contract. For most product groups, these three types of transactions account for well over 60 percent of all transactions. (See the Appendix for a description of the various types of transactions that comprise the sample.) One important point to note about these transactions is that an annual "contract" rarely means a price change every twelve months, nor does a monthly contract mean a

price change every month. Although annual contracts do involve more rigidity than monthly ones, it is incorrect to think of contracts as inflexible price rules set at specified intervals. A more appropriate view is that they are flexible agreements that can be renegotiated when and if the need arises.

The results in Table 2 show that, as one would expect from Table 1, the pattern of rigidity across product groups is highly varied. As a general rule, all product groups for each of the three transaction types in Table 2 are characterized by spells of price rigidity that in the majority of cases last less than one year. Some commodities like nonferrous metals and plywood are characterized by very flexible prices with over 60 percent of all spells in the monthly and quarterly monthly category lasting less than three months. On the other hand, there are definitely a substantial number of transactions involving very inflexible prices. For example, in steel, over 39 percent of the spells of rigid prices in the annual and quarterly monthly category (which comprises over half of all the transactions in steel) last more than one year. Other commodities with important transaction types showing fairly inflexible prices include paper, chemicals, cement and glass. In fact, a histogram analysis based on transactions (not spells) shows that 50 percent or more of all transactions involving steel, cement, chemicals, or glass, have average rigidities of one year or more for frequently used contract types.

As one would expect, the annual category involves less price flexibility than the quarterly category which itself exhibits less flexibility than the monthly category. It is also interesting to note that even within a particular product group and transaction type, there is a high degree of heterogeneity in price flexibility. For example, for chemicals monthly, over 50 percent of spells of rigidity are less than three months, but still a significant fraction (10 percent) involve spells of rigidity in excess of two years. This suggests that within any one product grouping, either the products sold are different, or the buyer-seller pairings have different properties, or the method chosen to allocate (i.e., price vs. nonprice) across different pairings of buyers and sellers is simply different.[4]

One issue frequently raised in discussions of price flexibility is the cost of making a price change (see, for example, Robert Barro, 1972). There are many types of costs associated with a price change. New price sheets have to be constructed, price information must be conveyed to buyers, buyers may find planning more difficult, buyers may distrust sellers if prices change often, search costs are higher if prices change often, and so on. The real question is how important are these costs. One way to address this question is to see how important small price changes are. Table 3 reports the percent of all price changes that are less than 1/4, 1/2, 1, and 2 percent, in absolute value for the same product groups and transaction types reported in Table 2.

Table 3 makes two points. First, very small price changes occur more often in monthly than in quarterly monthly or in annual transaction types. Second, and most important, there are a significant number of price changes that one would consider small (i.e., less than 1 percent) for most commodities and transaction types. This finding presents a bit of a puzzle if buyers are homogeneous. Either the cost of changing price is small or the costs of being at the "wrong" price—even one off by 1 percent—are very high.[5] Yet these explanations have difficulty explaining how it can be that some transactions seem to involve prices that do not change over long periods. Another explanation is that perhaps price does not need to

[4]Alternatively, the heterogeneity in spells could arise because supply and demand are changing over time. This last explanation turns out not to provide the full answer, as we shall see in Section VI. Moreover, a table analogous to Table 2, based on transactions, not spells, confirms the heterogeneity across transactions.

[5]Even if the fixed cost of changing price is small, one cannot necessarily rule out large welfare effects caused by this fixed cost. In a model with distortions, even small fixed costs can lead to large welfare losses. See, for example, N. Gregory Mankiw (1985) and George Akerlof and Janet Yellen (1985). Furthermore, the presence of even small fixed costs might affect the time-series properties of economic variables. See, for example, Julio Rotemberg (1982) and Olivier Blanchard (1982).

TABLE 3—FREQUENCY OF SMALL PRICE CHANGES BY PRODUCT GROUP BY CONTRACT TYPE

Product	Percent of Price Changes less than 1/4 Percent	1/2 Percent	1 Percent	2 Percent	Average Absolute Price Change (Percent)
Steel:					
Annual	4	8	11	27	3.3
Quarterly	5	11	17	24	4.2
Monthly	9	20	36	52	2.5
Nonferrous Metals:					
Annual	2	5	9	27	7.0
Quarterly	2	5	12	25	5.0
Monthly	8	15	28	49	2.9
Petroleum:					
Annual	0	0	8	24	5.3
Quarterly	0	0	2	17	5.4
Monthly	1	5	19	47	2.9
Rubber Tires:					
Annual	12	21	30	44	3.0
Quarterly	7	11	18	34	4.5
Monthly	13	23	38	63	2.3
Paper:					
Annual	4	9	8	27	6.3
Quarterly	0	19	24	33	3.6
Monthly	13	23	43	62	2.0
Chemicals:					
Annual	4	8	13	24	7.7
Quarterly	0	5	11	24	7.3
Monthly	5	14	30	42	5.0
Cement:					
Annual	14	22	32	46	3.3
Quarterly	0	0	1	19	4.1
Monthly	71	75	85	94	5.0
Glass:					
Annual	0	0	7	19	6.5
Quarterly	0	0	20	40	6.2
Monthly	3	20	45	67	2.1
Trucks, Motors:					
Annual	3	3	12	20	3.9
Quarterly	0	0	0	8	7.2
Monthly	12	27	50	75	1.7
Plywood:					
Annual	–	–	–	–	–
Quarterly	1	2	6	19	6.1
Monthly	19	38	54	72	1.9
Household Appliances:					
Annual	0	0	0	25	4.3
Quarterly	–	–	–	–	–
Monthly	22	44	70	95	.8

change in those transactions for which prices are unchanging (i.e., neither supply nor demand curves are shifting). This explanation runs into the problem that, as is suggested from Table 2 (and as will be confirmed later on), within the same product grouping there are likely to be changing prices for one transaction at the same time that there are constant prices for another. The only possible explanations consistent with efficiency seem to be either that firms differ in their allocation ability with some firms relying on

price more than others, or, alternatively, that every firm must rely more on price when dealing with certain buyers than with others.[6]

The foregoing analysis can also shed light on the question of whether there is an asymmetry in price movements. For example, are prices rigid downward? If prices are rigid downward, then one can think of the fixed cost of changing price as being higher for price declines than price increases. If so, the minimum positive price change should be less than the minimum negative price change. In fact, an analysis of minimum positive and negative price changes reveals no such pattern.

III. Relationship Between Price Rigidity, Price Change, and Length of Buyer-Seller Association

If within a particular product group, there is a wide degree of heterogeneity in price rigidity across buyers, are there any predictable correlations that emerge between price rigidity, price change, and length of buyer-seller association?[7] There are several different theories of price rigidity and the theories often have different implications for these correlations. I now investigate three questions.

First, is there a positive correlation between length of association and price rigidity across transactions for the same product?[8] That is, if buyer A has been dealing with his seller for ten years, while buyer B is beginning a new relationship, are buyer A's prices more rigid? One rationale for this relationship would be that if buyers and sellers deal with each other over long time periods, they set one average price and thereby save on the transaction cost of changing price constantly. However, it is quite possible to justify the reverse relationship. The impediment to changing price may be that the buyer or seller may feel the other side is taking advantage of him (see, for example, Williamson). If buyers and sellers have been dealing with each other for a long period of time, it will be in their interest not to take advantage of the other in the short run for fear of damaging the ongoing relationship (see, for example, Lester Telser, 1980). If buyers and sellers know each other well, because of their long-standing relationship, this fear of being taken advantage of in the short run will be reduced. In such a case, flexible prices may emerge.

Second, is there an inverse correlation between the size of price change and duration of price rigidity across transactions within the same product group? That is, if buyer A purchases steel on a contract in which price changes frequently, while buyer B has a contract in which price changes infrequently, are the price changes (when they occur) of buyer A larger (in absolute value) than those of buyer B? This relationship would make sense if prices are rigid on some transactions because there is a cost to changing price. If so, one would expect that those transactions with the most rigid prices (those to buyer B) have the highest costs of changing price and therefore only large price changes will be observed on those contracts. An alternative prediction would be that some prices are rigid because buyers (or sellers) want price stability for insurance-type reasons. In such a case, price changes on the more rigid con-

[6] I recognize the possibility that nonefficiency explanations may help explain some pricing behavior (for example, Akerlof-Yellen and Daniel Kahneman, Jack Knetsch, and Richard Thaler, 1986), but feel that the efficiency explanations have not yet been fully explored (see my forthcoming paper).

[7] Length of association is measured as the total time the buyer and seller have engaged in a transaction for a product whose specifications may change over the time of the association. This measure is a noisy one, because the buyer and seller may be engaged in other transactions which affect their knowledge of each other, and may have been dealing with each other prior to the beginning of the data set. Moreover, to the extent that a buyer reported prices from several suppliers, rather than one, for each reported price series, the measure of length of association is flawed. (See fn. 2.)

[8] See Table A1 for data by product group on average length of association and average price change. Table 1 reports average duration of price rigidity. Correlation of these three variables across product groups is not as good a way of uncovering systematic relationships among these three variables as is correlation of the three variables across transactions for the same product, because many factors differ between product groups.

tract could well be smaller than on the flexible price contract since the function of insurance is to smooth out price fluctuations.

The third question is whether there is a negative association between length of association and the size of price change. If buyers' and sellers' distrust of or lack of knowledge about each other explains rigid prices, then the longer the association, the lower the cost of changing price, and hence the more flexible should be price and the smaller the observed price changes. The opposite prediction could emerge from a theory in which buyers and sellers who deal with each other over long periods care about getting only the average price right. In such a case, one would expect to see rigid prices that infrequently change. When they do change, they will change by larger amounts than prices in less rigid contracts.[9]

Table 4 reports the correlations between length of association, price change (absolute value), and rigidity for each product group, and indicates when the correlations are statistically significant at the 10 percent level, 5 percent level, and 1 percent level.[10] A strong positive association between length of association and rigidity exists only for chemicals, while a strong negative association exists for petroleum, household appliances, and truck motors. To the extent any general relationship exists between length of association and rigidity, it is a *negative* one. The second column of Table 4 indicates that there is a *positive* association between price change and rigidity. All but one correlation is positive, and all seven statistically significant correlations are positive. The third column suggests that there is a *negative* correlation between length of association and price change. All but two correlations are negative, and all five statistically significant correlations are negative.

TABLE 4—CORRELATIONS OF CONTRACT CHARACTERISTICS

	Correlation Between:		
Product	Length of Association and Rigidity	Rigidity and Average Absolute Percent Price Change	Length of Association and Average Absolute Percent Price Change
Cement	.28	.17	.24
Chemicals	.16[c]	.10[a]	−.12[b]
Glass	−.11	.69[c]	−.24
Household Appliances[d]	−.87[c]	.71[b]	−.66[b]
Nonferrous Metals	.12	.12	−.15[b]
Paper	.03	.20	−.25[a]
Petroleum	−.25[c]	−.06	−.09
Plywood	.10	.54[c]	−.11
Rubber Tires	−.08	.43[c]	−.27[b]
Steel	.03	.14[b]	.01
Trucks, Motors	−.56[c]	.60[a]	−.23

[a]Statistical significance at the 10 percent level.
[b]Statistical significance at the 5 percent level.
[c]Statistical significance at the 1 percent level.
[d]Based on only 11 observations.

In short, the evidence in Table 4 is *consistent* with the following explanation. Buyers and sellers who do not have long associations are more likely to use fixed price contracts because they don't trust or know each other. The "cost" of changing price on such a contract is to risk creation of mutual distrust. Prices change on these contracts only for substantial price movements. Buyers and sellers who have long associations aren't as worried about mutual distrust. Hence, price changes are more frequent (i.e., less rigid prices) and on average smaller.[11]

One common explanation for price (or wage) rigidity has to do with insurance. I have not incorporated that explanation into the one just given for several reasons. First, recent work (Sherwin Rosen, 1985) casts doubt on the theoretical underpinnings of an

[9]This assumes that price changes are motivated by changes in the permanent price component whose changes are assumed larger than the transitory component. The reverse relation between permanent and transitory would flip the prediction.

[10]My 1986 working paper reports data on average rigidity, average price change, and average length of association by product by type of contract.

[11]A model that would generate such results would be one where costs are undergoing a random walk, production is constant returns to scale, and the cost of changing price is negatively related to length of association.

insurance explanation. Second, large firms should be able to diversify such risks, and hence not need insurance.[12] Third, as we will see in the next section, the insurance explanation does not seem supported by the data.

IV. Relationship Among Types of Transactions

Do some buyers seek out stable pricing arrangements in which the price changes infrequently? If so, one would expect to see a correlation in the rigidity of pricing across transactions of different commodities. For example, if the transactions of a particular buyer who purchased steel involved price changing much less frequently than the industry average, will it also be the case that the buyer's transactions involving paper have prices that change less frequently than the industry average?

For the product categories of Table 1, I have calculated for each buyer a vector of the average price rigidity for each of the commodities he purchases. I then examine pairs of products to see if there is a correlation across firms in these rigidities, (i.e., does a firm buying steel with overly rigid prices buy paper with overly rigid prices?) There are 227 buyer firms in my sample. There are many fewer (around 60) who purchase any two commodities. The pairwise correlations were primarily positive, but in most cases the correlations were not statistically significant, and were often sensitive to the interpolation method used to calculate price rigidity. The most stable and statistically significant results were the (positive) correlations between price rigidity for contracts in steel and rubber, metals and plywood, and rubber and cement.[13] Because of the instability of the results, these results should be regarded as at best weak support that buyers may have certain preferences across transaction types for different products.

V. Analysis of Specific Products

One drawback to the analysis of the previous sections is that the product groups may be so broad that a heterogeneity appears in the results which is caused only by the heterogeneous nature of the products in any one commodity group. To remedy this problem, I analyzed 32 specific products. These 32 products were chosen primarily because there were numerous data on them. The products analyzed are listed in Table 5 along with information similar to that presented in Table 1.

The results are similar to those of Table 1 in several respects. As in Table 1, there is wide variation across products in the rigidity of price. Even within a single detailed product specification, there still exists a great deal of heterogeneity in durations of spells of rigidity. The standard deviation of duration rises with the average duration.[14] One is struck by the rigidity of some prices. Even for monthly contracts, there are many products (for example, chlorine liquid, steel plate) where the average length of a spell of price rigidity is well over one year. And, column 6 indicates that, using transactions as the unit of analysis, most commodities have average durations of price in excess of eight months.

In Table 6, I present the histograms of spells of price rigidity by commodity for a frequently used contract specification. The pattern that emerges is similar to that in Table 2. Even within detailed product specification for a particular contract type, there is considerable heterogeneity in length of spells of price rigidity. This suggests that the price of a good is changing for some transactions but not for others.[15] Table 6 reveals that although most prices do not remain in effect for over one year, for many products (for example, steel plate, hot rolled bars and rods, oxygen) a significant number of spells (over 20 percent) of rigid prices remain in effect for over two years.

[12] This must be qualified by agency theories of monitoring.

[13] One curious finding is that price rigidity is negatively correlated at a statistically significant level for truck and steel contracts.

[14] The simple and rank correlations of average duration and the standard deviation of duration exceed .9.

[15] Histograms like Table 6 based on a transaction (not spell) as the unit of analysis confirm this.

TABLE 5—PRICE RIGIDITY FOR DETAILED PRODUCT SPECIFICATIONS[a]

Product (1)	Number of Buyer-Seller Pairings (2)	Average Duration Price Rigidity (Spells) (Months) (3)	Standard Deviation of Duration (Spells) (Months) (4)	Aver. Duration of Price Rigidity Monthly Contracts (Spells) (Months) (5)	Average Duration of Price Rigidity (Transactions) (Months) (6)
Steel plates	28	18.5	19.4	21.6	20.3
Hot rolled bars and rods	33	15.1	17.6	10.6	17.5
Steel pipe and tubing (3″ or less in diameter)	33	12.1	16.4	12.7	15.9
Copper wire and cable (bare)	26	3.8	5.4	2.6	4.1
Gasoline (regular)	66	6.2	5.7	2.7	8.9
Diesel oil #2	75	4.7	4.3	1.4	6.9
Fuel oil #2	41	7.3	4.9	4.6	8.3
Residual fuel oil #6	59	6.5	6.4	2.9	9.2
Container board, fiberboard	28	11.6	8.0	11.5	12.6
Caustic soda (liquid)	33	16.2	22.9	27.6	21.3
Chlorine liquid	28	19.9	18.7	60.0	27.1
Oxygen, cylinders	30	16.8	14.6	36.3	21.5
Acetylene	22	16.0	16.2	26.4	21.9
Portland cement (sack)	28	16.4	16.8	–	19.0
Steel sheet and strip, hot rolled	25	18.6	18.5	–	19.1
New rail (RR)	20	22.1	31.4	17.1	23.2
Tie plates (RR)	18	21.9	33.0	20.0	23.0
Steel wheels "one wear" (RR)	25	21.4	22.6	24.0	24.9
Track bolts (RR)	18	14.5	17.4	4.4	17.2
Zinc slab ingots	9	5.1	5.4	4.4	5.6
Coal (RR)	20	6.8	12.2	1.4	15.9
Kraft wrapping paper	12	7.5	6.0	5.7	9.2
Paper bags	16	9.4	5.3	20.0	10.3
Sulfuric acid, bulk	15	14.1	18.7	22.3	20.9
Sulfuric acid	19	11.0	17.1	5.1	19.5
Methyl alcohol	18	12.3	12.9	17.4	17.8
Phthalic anhydride	10	7.2	6.1	6.8	8.3
Succinate antibiotic	16	34.4	52.1	57.0	25.4
Kapseals antibiotic	16	56.1	66.7	40.0	44.0
Meprobanate tablets	16	13.8	12.0	18.7	15.5
Librium	13	19.1	23.1	56.0	20.9
Plywood	25	3.7	4.8	1.1	6.2

[a]See Table 1. The dashes in col. 5 indicate no data available.

In Table 7, I present the fraction of price changes that are less than 1/4, 1/2, 1, and 2 percent in absolute value in order to assess the importance of the fixed costs of changing price. Table 7 corroborates the message of Table 3. For most products, there are numerous (over 10 percent) instances of small price changes (below 1 percent). This fact reinforces my earlier conclusion that theories that postulate rigid prices solely because of a common high fixed cost of changing price to each buyer are not supported by the evidence. (See the discussion of the results of Table 3.) The most reasonable explanation is that firms and buyers must differ in their need to rely on the price system to achieve

TABLE 6—HISTOGRAMS OF DURATIONS OF RIGIDITY BY DETAILED PRODUCT SPECIFICATION BASED ON SPELLS OF RIGIDITY

Product	0–3 Mo.	3 Mo.– 1 Yr	1–2 Yrs	2–4 Yrs	Over 4 Yrs
Steel plate	.24	.24	.23	.18	.11
Hot rolled bars and rods	.36	.21	.21	.16	.07
Steel pipe and tubing (less than 3″ diameter)	.39	.31	.16	.10	.05
Copper wire and cable (bare)	.67	.30	.02	0	.01
Gasoline (regular)(A)	.33	.59	.05	.03	0
Diesel oil #2	.79	.22	0	0	0
Fuel oil #2 (A)	.03	.88	.08	.02	0
Residual fuel oil #6 (A)	.22	.64	.07	.06	0
Container board, fiberboard (A)	0	.73	.19	.06	0
Caustic soda (liquid)(A)	.10	.64	.14	.06	.06
Chlorine liquid (A)	0	.69	.14	.10	.06
Oxygen, cylinders	.32	.27	.14	.26	.01
Acetylene	.37	.24	.15	.21	.01
Portland cement (bag or sack)	.19	.32	.24	.14	.05
Steel sheet and strip, hot rolled	.25	.27	.19	.21	.08
New rail	.53	.07	.16	.06	.18
Tie plates	.53	.08	.17	.06	.16
Steel wheels "one wear"	.13	.35	.22	.22	.09
Track bolts	.27	.34	.23	.06	.11
Zinc slab ingots	.44	.44	.09	.03	0
Coal, for RR	.60	.23	.11	.03	.03
Kraft wrapping paper	0	.40	.40	.20	0
Paper bags	.17	0	.67	.17	0
Sulfuric acid, bulk	.68	.18	.08	0	.05
Sulfuric acid (A)	.13	.56	.20	.05	.05
Methyl alcohol (A)	.38	.38	.15	.07	.01
Phthalic anhydride	.47	.41	.09	.03	0
Succinate antibiotic	0	.30	0	.50	.20
Kapseals antibiotic	0	.08	.08	.31	.54
Meprobanate tablets (A)	.14	.67	.11	.06	.03
Librium (A)	.13	.39	.22	.17	.09
Plywood	.73	.23	.03	.01	.01

Note: All contracts are monthly or quarterly monthly, unless followed by (A) which indicates annual. The numbers in rows may not add to one because of rounding.

allocative efficiency and that the fixed costs of changing price varies across buyers and across firms.

An analysis of the minimum positive and negative price changes reveals no tendency for one to exceed the other. Just as in the earlier analysis, there appears to be no evidence to support asymmetric price changes.

In Table 8, I present information, comparable to Table 4, on the relationship between price rigidity, length of association, and average price change for transactions in the same product.[16] (Table A2 in the Appendix presents information by product on average absolute price change and average length of association.) The results mirror those of Table 4. There may be a weak negative correlation between rigidity and length of association. Of the 18 correlations,

[16]Most correlations involve between 20 to 30 observations, with 15 being the minimum number of observations required in order to be reported.

TABLE 7—FREQUENCY OF SMALL PRICE CHANGES BY DETAILED PRODUCT SPECIFICATION

	Percent of Price Changes less than			
Product	1/4 Percent	1/2 Percent	1 Percent	2 Percent
Steel plate	0	1	11	16
Hot rolled bars and rods	1	8	13	28
Steel pipe and tubing (less than 3″ diameter)	4	6	14	27
Copper wire and cable (bare)	3	5	8	19
Gasoline (regular)(A)	0	1	13	27
Diesel oil #2	0	0	2	19
Fuel oil #2 (A)	0	0	7	22
Residual fuel oil #6 (A)	0	0	2	25
Container board, fiberboard (A)	4	4	4	12
Caustic soda (liquid)(A)	2	5	11	15
Chlorine liquid (A)	6	13	17	31
Oxygen, cylinders	0	0	3	14
Acetylene	0	10	18	23
Portland cement (bag or sack)	0	0	1	19
Steel sheet and strip, hot rolled	0	2	7	13
New rail	1	3	6	10
Tie plates	3	5	5	9
Steel wheels "one wear"	4	4	10	16
Track bolts	1	3	14	16
Zinc slab ingots	6	6	11	20
Coal (RR)	3	8	18	37
Kraft wrapping paper	3	8	20	53
Paper bags	0	20	20	60
Sulfuric acid, bulk	3	12	34	54
Sulfuric acid	1	1	57	76
Methyl alcohol (A)	0	15	24	32
Phthalic anhydride	0	0	0	0
Succinate antibiotic	0	0	0	0
Kapseals antibiotic (A)	0	0	0	50
Meprobanate tablets (A)	0	0	0	27
Librium (A)	0	0	0	14
Plywood	1	3	7	18

Note: See Table 6.

only 4 were statistically significant. Two negative correlations were significant at the 1 percent level, while the positive correlations were significant at the 5 and 10 percent levels. (However, the number of positive correlations exceeded the number of negative ones.) The evidence on the correlation between price change and rigidity is clearer. Of the 9 significant correlations, 8 were positive. The number of positive correlations exceeded the number of negative ones. The evidence on the correlation between price change and length of association suggests a negative correlation. Of the 5 significant coefficients, all were negative. (However, the number of negative correlations equalled the number of positives.)

VI. The Heterogeneity of Price Movements Across Buyers

The previous evidence reveals that price movements across different transaction types for the same commodity may be very differ-

TABLE 8—CORRELATIONS OF CONTRACT CHARACTERISTICS

Product	Correlation Between: Length of Association and Rigidity	Rigidity and Average Absolute Percent Price Change	Length of Association and Average Absolute Percent Price Change
Steel sheet and strip, hot rolled	–	−.40[a]	–
Steel plate	.07	−.11	.27
Hot rolled bars and rods	−.00	.32[a]	.26
Steel pipe and tubing (3″ or less in diameter)	−.21	.19	−.32[a]
Plywood	.10	.04	−.34[a]
New rail	.14	.41[a]	−.64[b]
Tie plates	–	.47[b]	–
Steel wheels "One wear"	.07	−.33	−.14
Track bolts	–	.54[b]	–
Copper wire and cable, bare	−.06	.76[c]	−.20
Coal, for RR	–	−.14	–
Gasoline	.02	.08	−.02
Diesel oil #2	−.74[c]	−.22	.27
Fuel oil #6	−.12	−.02	−.14
Sulfuric acid, bulk	.51[b]	−.06	−.45[b]
Sulfuric acid	−.52[c]	.15	.10
Caustic soda, liquid	.35	.58[c]	.22
Chlorine, liquid	.40[a]	−.00	−.56[b]
Oxygen cylinders	.10	−.17	.07
Acetylene	.04	.50[b]	.12
Methyl alcohol	.21	.53[c]	.02
Portland cement, in bag or sack	.34	.19	.33

[a]Significance at the 10 percent level.
[b]Significance at the 5 percent level.
[c]Significance at the 1 percent level.

ent. In this section, I investigate in more detail the heterogeneity of price movements for the same commodity. By limiting the analysis to transactions of the same type, I have automatically screened out considerable heterogeneity. Despite this, I still find a startling amount of heterogeneity. I limit the analysis to annual contracts or quarterly monthly and monthly contracts, depending on the available data. I group price movements from quarterly monthly and monthly together on the grounds that they both represent price series whose prices are not necessarily expected to remain in force for more than one month.

I use two methods to describe how heterogeneous price movements are. The first method measures the difference in the stochastic structure of each price change series while the second attempts to measure correlation in price movements across different transactions.

The first method computes for each individual price series the variance in the percent changes in price (actually the first difference of the log of the price series). A variance σ is computed for each transaction price series. If all the price series have the same stochastic structure, this variance should be the same across different price series for the same commodity. For each of 30 commodities, I present the mean variance (i.e., the mean of σ^2), the variance of σ^2 (i.e., a measure of how σ^2 varies across transactions), and the coefficient of variation (square root of variance of σ^2 divided by the mean).[17]

Table 9 shows that, in general, the individual price series within any one commodity and transaction type seems to be quite different from one another. The commodities that seem to have the least homogeneous transactions are steel pipe, oxygen, sheet steel, steel railway wheels, and coal.

Another method of characterizing the degree of heterogeneity among price series is to look at the correlation of contemporaneous price changes. A slight extension of this method is to examine the correlation of filtered price series. An example will illustrate.

Suppose two monthly price series are

	10	10	10	10	5	5	5	5	7.5	7.5	7.5	7.5,
and	10	10	10	5	5	5	5	7.5	7.5	7.5	7.5	7.5

One might be especially interested in seeing how closely the percent changes in the price series are correlated. The two derived series of percent price changes are

–	0	0	0	−50%	0	0	0	50%	0	0	0
–	0	0	−50%	0	0	0	50%	0	0	0	0

[17]Some products from Table 5 were dropped because of data incompleteness.

TABLE 9—MEASURES OF HETEROGENEITY AMONG PRICE SERIES

Product	Mean Variance of Individual Price Change	Variances of Individual Price Change	Coefficient of Variation
Steel plate	1.33 (10–6)	1.56 (10–9)	29.7
Hot rolled bars and rods	1.73 (10–6)	3.64 (10–9)	34.9
Steel pipe and tubing (3″ or less in diameter)	3.31 (10–6)	2.27 (10–8)	45.5
Copper wire and cable, bare	1.45 (10–5)	4.36 (10–8)	14.4
Gasoline	6.22 (10–5)	1.03 (10–6)	16.3
Diesel #2	1.59 (10–5)	6.50 (10–8)	16.0
Fuel oil #2 (A)	2.93 (10–5)	1.02 (10–7)	10.9
Fuel oil #6	2.57 (10–5)	4.54 (10–7)	26.2
Container board, fiberboard	2.94 (10–5)	5.62 (10–9)	2.5
Caustic soda, liquid	5.26 (10–5)	4.89 (10–8)	4.2
Liquid chlorine (A)	8.48 (10–6)	6.57 (10–8)	30.2
Oxygen, cylinders	3.07 (10–5)	2.49 (10–6)	51.4
Acetylene	6.66 (10–6)	4.63 (10–8)	32.3
Portland cement	1.97 (10–6)	4.79 (10–9)	35.1
Sheet steel and strip (hot rolled)	4.64 (10–6)	1.63 (10–7)	87.0
New rails	9.95 (10–7)	1.44 (10–10)	12.1
Tie plates	1.55 (10–6)	1.43 (10–10)	7.7
Steel railway wheels	9.51 (10–7)	8.08 (10–9)	94.5
Railroad track bolts	2.87 (10–6)	4.93 (10–9)	24.5
Zinc slab, ingot	6.21 (10–5)	7.09 (10–8)	4.3
Coal (RR)	9.15 (10–6)	1.60 (10–7)	43.7
Sulfuric acid, bulk (A)	5.92 (10–5)	1.91 (10–6)	23.3
Sulfuric acid (A)	5.54 (10–5)	9.05 (10–7)	17.2
Methyl alcohol (A)	7.24 (10–5)	1.55 (10–7)	5.4
Phthalic anhydride	2.78 (10–4)	1.52 (10–6)	4.4
Succinate (A)	5.42 (10–6)	3.13 (10–8)	32.6
Kapseals (A)	2.52 (10–6)	2.77 (10–9)	20.9
Meprobanate tablets (A)	2.59 (10–4)	3.83 (10–6)	7.6
Librium (A)	6.39 (10–5)	5.40 (10–7)	11.5
Plywood	2.08 (10–5)	1.43 (10–7)	18.2

Note: See Table 6.

It appears that the two series have no correlation in percent changes. But that conclusion is misleading. Both series change within one month of each other. Suppose that one constructs a new series that takes the arithmetic average of the last two monthly percent changes in prices. Then one obtains two series that look like

–	–	0	0	–25%	–25%	0	0	25%	25%	0	0
–	–	0	–25%	–25%	0	0	25%	25%	0	0	0

The correlation between the two new series will be positive and will equal .5. If one uses a three-month filter (i.e., average over the last three monthly percent changes in price), the correlation rises to .67. In general, one initially expects correlation to rise as the period of averaging increases.

Before presenting tabulations of correlations by product for different filter sizes, it will be helpful first to decide what is a "high" or "low" correlation. In other words, we must develop some underlying standard as to how closely two very related series should move. Suppose we adopt the position that two price series that change by identical amounts within, say, three months of each other are "highly" correlated. Let $\rho(F)$ be

the contemporaneous correlation of the two price series when averaging over F periods is performed. Suppose that the two series representing percent price changes are identical, are displaced from each other by three months, and that price changes are independent of the preceding price change. Then, it is easy to show that

$$\rho(1)=\rho(2)=0$$

$$\rho(F)=1-3/F \qquad F>3.$$

This means that for a filter of size 6, the correlation between our two series is .5, and rises to .75 for filters of one year. In general, one should expect that very high correlations (above .7) will probably be unusual for filters below twelve months, even for "well-behaved" price series.

Each of 30 products was analyzed separately. For each product, and for each contract type an average correlation for a particular filter size was computed. For example, suppose that there are 10 individual contract transactions for steel plates, each lasting ten years. The monthly percent change in price (differences in log of price) was calculated for each series for each month. The simple correlation was computed for every combination of contracts (i.e., 45 pairs) and an average correlation over the 45 pairs was then computed. If the average correlation is high, it says that on average the price series move together. If the average correlation is low, it suggests that price movements for the same good are only very loosely related to each other. If the low correlation persists as the filter increases to say two years, it says that knowing how person A's price has changed over a two-year period doesn't help much in predicting how person B's price will change (averaged over the two-year period).

In Table 10, I present measures of average correlation for filters of one month and twelve months for each of the 30 commodities for selected contract types.[18] As expected, $\rho(12)$ usually exceeds $\rho(1)$. If we use the criterion that correlations on the order of .5 and above represent price series that move pretty closely together, we see that for several

TABLE 10—HETEROGENEITY MEASURES: CORRELATIONS AMONG PRICE SERIES

Product[a]	$\rho(1)$[b]	$\rho(12)$[b]
Steel plate (M)	.42	.61
Hot rolled bars and rods (M)	.42	.60
Steel pipe and tubing (M) (3″ or less in diameter)	.16	.25
Copper wire and cable (M)	.53	.78
Gasoline (A)	.02	.07
(M)	.04	.30
Diesel fuel #2 (A)	.00	.06
(M)	.53	.69
Fuel oil #2 (A)	.01	−.03
Fuel oil #6 (A)	.02	.11
(M)	.26	.49
Container board, fiberboard (A)	.14	−.03
(M)	.06	.16
Caustic soda, liquid (A)	.07	.07
(M)	.04	.36
Liquid chlorine (A)	.05	.08
Oxygen, cylinders (A)	.03	.17
(M)	.28	.40
Acetylene (M)	.30	.54
Portland cement (M)	.15	.21
Steel sheet and strip, (M) hot rolled	.40	.44
Rails (M)	.81	.94
Tie plates (M)	.78	.88
Steel railway wheels	.37	.54
Railroad track bolts	.47	.62
Zinc slab ingots (M)	.52	.90
Coal (RR) (M)	.14	.17
Phthalic anhydride (M)	.27	.68
Sulfuric acid, bulk (A)	.13	.32
Sulfuric acid (A)	.10	.07
Methyl alcohol (A)	.22	.46
Succinate (A)	0.0[c]	0.0[c]
Kapseals (A)	0.0[c]	0.0[c]
Meprobanate tablets (A)	.03	−.07
Librium (A)	−.02	−.06
Plywood (M)	.16	.21

[a] Contracts are either quarterly monthly or monthly (indicated by M) or annual (indicated by A).
[b] $\rho(i)$: Correlations of price changes averaged over i months.
[c] No price movement in most contracts.

[18] Filters of 2 years produced results similar to those for filters of 1 year. Correlations were also calculated on the timing of price changes (i.e., 0 or ± 1 indicating whether or not a price change occurred and its direction) and the same low correlations persisted.

products, there is a homogeneity of price movements. On the other hand, there are several products like cement, container board, plywood, and several chemical products that have very low (sometimes even negative) correlations even for twelve-month averaging. In fact, it is startling to find so many products where it is clear that some mechanism other than only price is allocating resources.[19] It is noteworthy that container board exhibits low correlations of price, since I understand that quantity rationing is sometimes used in the paper industry in place of price rationing.[20]

It is interesting to see whether there is any agreement between the two methods of characterizing heterogeneity in Tables 9 and 10. In fact, there is a low degree of agreement. The simple correlation between the measures of heterogeneity in Tables 9 and 10 is below .1 and is not statistically significant. On the other hand, there is a high degree of statistically significant (negative) correlation between $\rho(1)$ (or $\rho(12)$) and other measures of heterogeneity such as the coefficients of variation for rigidity, price change, and length of association.[21] This may imply that the measure in Table 9 is capturing an aspect of price different from the other measures or, alternatively, that the measure in Table 9 is not a useful one.

VII. Implications for Price Behavior

Tables 1 through 10 can form the foundation for several predictions. For example, one could predict the following:

1) The products with high correlations for $\rho(12)$ in Table 10 should tend to have more serial correlation in their *WPI* component than products with low correlations;

2) Industrywide price adjustment for products with high values for $\rho(1)$ in Table 10 should tend to be swift;

3) Price controls on products with long spells of rigid prices (Table 1) are less likely to have harmful efficiency effects than controls on products with short spells of rigidity because nonprice methods are probably already used for products with very rigid prices to allocate resources.

I have not systematically investigated these three claims for each of the products listed in Table 10. However, I have done some work to corroborate at least some of the claims for some products. For example, from Table 10 copper wire and cable has a $\rho(12)$ of .78 while gasoline (monthly) has a $\rho(12)$ of only .30. The correlation between the monthly *WPI* and the monthly *WPI* lagged once (1957–66) for copper wire and cable is .99 which, as expected, exceeds that same measure (.88) for gasoline.[22]

Michael Bordo (1980) has estimated adjustment lags in prices for some of the commodity groups well represented in Table 10, such as metals and metal products, chemicals, and fuel. Based on the size of $\rho(1)$ in Table 10, I would predict the speed of adjustment to be fastest in metals and metal products, and the speed of adjustment in fuels and chemicals to be much slower and roughly equal to each other. In fact, Bordo (p. 1105) finds the mean lag of price adjustment for metals and metal products to 3.66 months, while the lag for fuels and chemicals are 6.64 and 6.20 months, respectively.

Finally, the only evidence I could find on the difficulty of price controls is John Kenneth Galbraith's *A Theory of Price Control* (1952) which is an account of his experience in controlling prices during World War II when he headed the Office of Price Administration (OPA). Although he does not deal explicitly with all the products in Tables 1–10, he does talk about steel products, which from Table 1 have a high degree of price rigidity. Galbraith states: "The Office of Price Administration controlled the price of all steel mill products with far less manpower and trouble than was required for

[19] My 1979 article presents a theory on buyer heterogeneity, which shows how prices to different buyers can exhibit low (or negative) correlations.

[20] Based on personal discussions with industry members.

[21] Table A3 in the Appendix reports these correlations.

[22] The source for *WPI* data was Stigler and Kindahl (Appendix C).

a far smaller volume of steel scrap...it is relatively easy to fix prices that are already fixed" (p. 17).

Although bits of evidence corroborate the predictions for some types of commodities, they obviously are far from conclusive. They do, however, show the value of evidence like that in Tables 1 through 10.

VIII. Structural Determinants of Price Behavior

Is there any correlation between industry characteristics and any of the measures of heterogeneity such as those in Tables 9 and 10? Using 30 products, I correlated the measures of heterogeneity in price movements of Tables 9 and 10 with the following variables: 1) mean absolute growth and variability of price (the higher is this number the higher the expected correlation of price movements); 2) measures of competitiveness (a) four-firm concentration ratio and (b) fraction of shipments beyond 500 miles; 3) growth and variability of total industry shipments; 4) length of buyer-seller association.

Simple correlations never emerged statistically significant (with the exception of the variance of the growth rate in price), though the correlations were generally in the positive direction. However, since no more than 30 observations were available, it would be premature to conclude that these structural characteristics do not influence price heterogeneity in the industry.

Is there any correlation between concentration and duration of price rigidity or length of association or average price change? The only significant correlation was between concentration (four firms) and duration of price rigidity. That correlation was statistically significant at the 5 percent level. The correlation implies that for every 10 point increase in the four-firm concentration ratio, prices remain rigid for an extra 1.6 months.[23] This finding is particularly interesting because it suggests that allocations are performed differently in concentrated and unconcentrated markets. I believe it is premature to draw the conclusions, implicit in the work of Means, Arthur Burns (1936), Galbraith, and others, that the markets have stopped working when they become concentrated. Instead, an alternative interpretation is that as firms become large they supplant the market's exclusive reliance on price as an allocation device and resort to other methods. In a world filled with transaction costs, exclusive reliance on a market-generated price to allocate goods could well be inferior to other nonprice allocation methods. It is the case, however, that markets that use nonprice allocation will respond to market shocks much differently than markets that exclusively use price to allocate. See my forthcoming paper for a fuller development of this theory.

IX. Conclusions

Since this paper began with a summary of the empirical results, I will not repeat them here. The main conclusion is that several of the empirical results are sufficiently startling that we should reexamine the central, often exclusive, role assigned to the price mechanism in theories of efficient resource allocation. It is not necessarily that the price mechanism has failed, but rather that alternative allocation mechanisms are used in addition to the price mechanism to achieve efficiency.

[23] The *OLS* equation is (standard errors in parentheses)

$$Av.\ Duration = \underset{(3.12)}{4.97} + \underset{(6.08)}{16.12}\ CR\,4 \quad \underset{SEE = 4.9}{R^2 = .22}$$

where *Av. Duration* is the average length of a spell of price rigidity and *CR 4* is the four-firm concentration ratio. This equation is based on 27 observations. The *CR 4* variable is the 1963 four-firm concentration ratio for the 5-digit SIC code that seems to correspond to the product. This correspondence is not exact and, for that reason together with the small number of observations, the results should be regarded with some caution. Another interesting finding involving concentration is that concentrated industries have a greater frequency of small price changes.

TABLE A1—CHARACTERISTICS OF CONTRACTS BY PRODUCT

Product	Average Length of Association Between Buyer and Seller (Months)	Average Size of Absolute Value of Percent Price Change	Product	Average Length of Association Between Buyer and Seller (Months)	Average Size of Absolute Value of Percent Price Change
Steel	105	3.5	Cement	103	3.0
Nonferrous Metals	86	4.0	Glass	91	4.2
Petroleum	87	4.4	Truck Motors	82	2.7
Rubber Tires	98	3.9	Plywood	114	5.0
Paper	91	3.4	Household Appliances	75	1.0
Chemicals	81	7.0			

TABLE A2—CHARACTERISTICS OF CONTRACTS BY PRODUCT

Product	Average Length of Association Between Buyer/Seller (Months)	Average Size of Absolute Value of Percent Price Change	Product	Average Length of Association Between Buyer/Seller (Months)	Average Size of Absolute Value of Percent Price Change
Steel plates	108	3.8	New rail (RR)	116	3.9
Hot rolled bars and rods	109	3.7	Tiel plates (RR)	119	4.5
Steel pipe and tubing (3″ or less in diameter)	114	4.6	Steel wheels "one wear" (RR)	119	3.8
Copper wire and cable (bare)	68	4.4	Track bolts (RR)	119	4.2
			Zinc slab ingots	104	4.8
Gasoline (regular)	91	3.3	Coal (RR)	119	3.7
Diesel oil #2	94	4.3	Kraft wrapping paper	94	4.3
Fuel oil #2	89	4.6	Paper bags	88	4.8
Residual fuel oil #6	73	5.8	Sulfuric acid, bulk	96	4.8
Container board, fiberboard	78	5.2	Sulfuric acid	103	3.5
			Methyl alcohol	91	7.1
Caustic soda (liquid)	84	7.8	Phthalic anhydride	93	11.7
Chlorine liquid	89	5.0	Succinate antibiotic	58	8.3
Oxygen, cylinders	109	11.5	Kapseals antibiotic	70	14.9
Acetylene	116	6.9	Meprobanate tablets	64	12.1
Portland cement (by sack)	104	3.7	Librium	48	8.6
			Plywood	110	5.2
Steel sheet and strip, hot rolled	120	5.9			

TABLE A3—CORRELATIONS AMONG MEASURES OF HETEROGENEITY

	CV DUR	*CV DP*	*CV ASSOC*	*CV VAR*	$\rho(1)$	$\rho(12)$
CV DUR	1	.88[a]	.41[a]	−.03	−.63[a]	−.60[a]
CV DP		1	.35[a]	.39[a]	−.57[a]	−.66[a]
CV ASSOC.			1	−.58[a]	−.47[a]	−.30
CV VAR				1	.08	−.01
$\rho(1)$					1	.91[a]
$\rho(12)$						1

Notes: CV DUR = coefficient of variation of duration; *CV DP* = coefficient of variation of the absolute value of price change (log difference); *CV ASSOC* = coefficient of variation of the length of association; *CV VAR* = coefficient of variation of the actual price changes counting no change as zero change; $\rho(1), \rho(12)$ = correlations of price changes averaged over i months.

[a]Significant at 5 percent level.

APPENDIX

Transactions were classified into one of ten categories by Stigler and Kindahl. The most important classifications include:

Annual contract: contract in force for one year.
Annual average: average of transaction prices during the year.
Annual monthly: annual observations of a transaction that occurs monthly.
Semiannual contract: contract in force for six months.
Semiannual average: average of transaction prices during six months.
Quarterly contract: contract in force for three months.
Quarterly average: average of transaction prices during the quarter.
Quarterly monthly: quarterly observation of a transaction that occurs monthly.
Irregular: irregular.
Monthly: monthly observations of a transaction that occurs monthly.

Tables 1A and 2A of my 1986 working paper report the importance of each classification by product group and for individual products.

REFERENCES

Akerlof, George A. and Yellen, Janet L., "Can Small Deviations from Rationality Make Significant Differences to Economic Equilibrium?," *American Economic Review*, September 1985, *75*, 708–20.

Barro, Robert, "A Theory of Monopolistic Price Adjustment," *Review of Economic Studies*, January 1972, *39*, 17–26.

Blanchard, Olivier J., "Price Desynchronization and Price Level Inertia," NBER Working Paper 900, 1982.

Bordo, Michael, "The Effects of Monetary Change on Relative Commodity Prices and the Role of Long-Term Contracts," *Journal of Political Economy*, December 1980, *88*, 1088–109.

Burns, Arthur, *The Decline of Competition*, New York: McGraw-Hill, 1936.

Carlton, Dennis W., The Rigidity of Prices, NBER Working Paper 1813, 1986.

______, "The Theory and The Facts of How Markets Clear: Is Industrial Organization Valuable for Understanding Macroeconomics?," in R. Schmalensee and R. Willig, eds., *Handbook of Industrial Organization*, Amsterdam: North-Holland, forthcoming.

______, "Equilibrium Fluctuations When Price and Delivery Lags Clear the Market," *Bell Journal of Economics*, Autumn 1983, *14*, 562–72.

______, "The Disruptive Effect of Inflation on the Organization of Markets," in R. E. Hall, ed., *Inflation*, Chicago: University of Chicago Press, 1982.

______, "Contracts, Price Rigidity, and Market Equilibrium," *Journal of Political Economy*, October 1979, *87*, 1034–62.

______, "Market Behavior with Demand Uncertainty and Price Inflexibility," *American Economic Review*, September 1978, *68*, 571–87.

Galbraith, J. K., *A Theory of Price Control*, Cambridge: Harvard University Press, 1952.

Hall, Robert E., "The Apparent Rigidity of Prices," NBER Working Paper 1347, 1984.

Kahneman, Daniel, Knetch, Jack L. and Thaler, Richard, "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review*, September 1986, *76*, 728–41.

Mankiw, N. Gregory, "Small Menu Costs and Large Business Cycles: A Macroeconomic Model of Monopoly," *Quarterly Journal of Economics*, May 1985, *100*, 529–38.

Means, Gardiner C., "Industrial Prices and Their Relative Inflexibility," U.S. Senate Document 13, 74th Congress, 1st Session, Washington 1935.

Mills, Frederick C., *The Behavior of Prices*, NBER General Series, No. 11, New York: Arno Press, 1927.

Okun, Arthur, *Prices and Quantities: A Macroeconomic Analysis*, Washington: The Brookings Institution, 1981.

Rosen, Sherwin, "Implicit Contracts: A Survey," *Journal of Economic Literature*, September 1985, *23*, 1144–75.

Rotemberg, Julio "Monopolistic Price Adjust-

ment and Aggregate Output," *Review of Economic Studies*, October 1982, *49*, 517–31.

Stigler, George and Kindahl, James, *The Behavior of Industrial Prices*, NBER General Series, No. 90, New York: Columbia University Press, 1970.

Telser, Lester, "A Theory of Self-Enforcing Agreements," *Journal of Business*, January 1980, *53*, 27–44.

Williamson, Oliver, *Markets & Hierarchies Analysis and Antitrust Implications: A Study in the Economies of Internal Organization*, New York: Free Press, 1975.

A Tax-Based Test for Nominal Rigidities

By James M. Poterba, Julio J. Rotemberg, and Lawrence H. Summers*

In macroeconomic models with flexible wages and prices, whether a tax is levied on producers or consumers does not affect its ultimate incidence. This equivalence breaks down in the presence of short-run nominal rigidities. Using both British and American data, we provide evidence against complete wage and price flexibility.

The side of a market on which a tax is levied is irrelevant in the standard microeconomic analysis of taxation. Students in elementary economics learn that it makes no difference whether a sales tax is collected from buyers or sellers. They are taught that the ultimate incidence of a payroll tax depends on the elasticities of supply and demand for labor, not on whether the tax is levied on employees or employers. Broader equivalence results concerning sales and income taxes are central to the analysis of general equilibrium tax incidence. Standard Keynesian macroeconomic analyses take a very different view. Raising sales taxes is thought to be inflationary, even if monetary policy remains unchanged. There is less concern that increases in direct taxation will increase prices.

The microeconomic and Keynesian views diverge because the former presumes flexible wages and prices, while the latter postulates rigid nominal wages. With wage rates fixed in the short run, sales tax increases necessarily raise prices; income tax increases have no such effect. If nominal wages are rigid over reasonable lengths of time, then the conventional tax analysis must be altered. Holding monetary policy constant, increases in the price level translate point for point into reductions in output. Even a temporary 1 percent decline in *GNP* could dwarf the potential efficiency gains from many proposed tax reforms.

The very existence of nominal rigidities is a subject of contemporary macroeconomic debate. Many Keynesian scholars take it as self-evident that nominal wages are sticky, at least in the short run. For example, Robert Solow invites his readers to "accept the apparent evidence of one's senses and [take] it for granted that the wage does not move flexibly to clear the labor market" (1980, p. 8). Other researchers claim that there is no available evidence in support of this hypothesis. For example, Robert King and Charles Plosser write that "Keynesian models typically rely on implausible wage or price rigidities, from the textbook reliance on exogenous values to the recent more sophisticated effort of Fischer that relies on nominal contracts" (1984, p. 363).

Examining how changes in the money stock affect macroeconomic activity, a standard test for nominal rigidities, is unlikely to resolve this issue conclusively. Shifts between direct and indirect taxation can provide tests which avoid many of the difficulties with money-based tests, since they are less likely to be endogenous responses to macroeconomic events. This paper employs

*Department of Economics, MIT, Cambridge, MA 02139; Alfred P. Sloan School of Management, MIT, Cambridge, MA 02139; and Department of Economics, Harvard University, Cambridge, MA 02138, respectively. We are indebted to Craig Alexander and Ignacio Mas for excellent research assistance, to Olivier Blanchard, Rudiger Dornbusch, Stanley Fischer, Greg Mankiw, two referees, and the co-editor, John Taylor, for helpful suggestions, and to the National Science Foundation for financial support. Frank DeLeeuw of the U.S. Bureau of Economic Analysis, and R. Doggett, K. Newman, and A. Tansley of the U.K. Central Statistical Office provided us with unpublished data. The research reported here is part of the NBER's research programs in Economic Fluctuations and Taxation. Any opinions expressed are our own and not those of the NBER or NSF.

both British and American data to investigate how shifts in the direct vs. indirect tax mix affect wages, prices, and output. Our results support the existence of nominal rigidities and suggest that they may have important effects which should be recognized when analyzing the short-run effects of tax reform.

The paper is divided into five sections. Section I clarifies the equivalence of direct and indirect taxation when wages and prices are fully flexible. It also shows how these equivalences fail when nominal rigidities are introduced. Section II describes our methodology for examining the impact of tax changes. The next section explains how we construct effective direct and indirect tax rates for Britain and the United States. Section IV presents our empirical findings based on postwar time-series evidence from both countries. The concluding section sketches the implications of our results for the analysis of tax policy and macroeconomic fluctuations.

I. Shifts from Direct to Indirect Taxation: Classical and Keynesian Views

In textbook public finance models, the legal incidence of a sales tax is of no consequence. It does not matter whether the tax is collected from producers or consumers. More generally, the equivalence theorems summarized in George Break (1974) establish that in an economy without savings and with flexible prices, a sales tax on all goods is equivalent to an equal-revenue tax on all income. This section presents a simple classical macroeconomic model illustrating these results, and shows how they break down when wage and price stickiness is introduced.

The equivalence between sales and income taxation is easily demonstrated with perfectly flexible wages and prices.[1] In the short run, aggregate output (Y) is a function only of labor input (N):

$$Y = f(N). \tag{1}$$

With competitive firms, an aggregate labor demand schedule can be obtained by equating the marginal product of labor to the firm's real wage:

$$f'(N) = w/s, \tag{2}$$

where w is the nominal wage and s is an index of prices received by firms. The supply of labor depends on the purchasing power of the worker's payment for an hour of work:

$$N = N\left[w(1-\tau)/s(1+\theta)\right], \tag{3}$$

where τ is the income tax rate and θ is the sales or, equivalently, value-added tax rate. Labor supply is unaffected by any reform which does not change $(1-\tau)/(1+\theta)$.

The government raises revenue from both income and sales taxes. Tax collections, T, equal

$$T = \tau Y + \theta E, \tag{4}$$

where E is household expenditure on goods and services, net of sales taxes. The household budget constraint in our one-period model is

$$(1-\tau)Y = (1+\theta)E. \tag{5}$$

Total revenue is then

$$\begin{aligned} T &= \left[\tau + \theta((1-\tau)/(1+\theta))\right]Y \\ &= \left[1-(1-\tau)/(1+\theta)\right]Y. \end{aligned} \tag{6}$$

For a given tax base, tax revenue depends only on $(1-\tau)/(1+\theta)$.

Consider the effects of a balanced budget tax reform which increases θ and reduces τ while leaving $(1-\tau)/(1+\theta)$ constant.[2] Both

[1] An appendix available from the authors on request shows that the equivalence between direct and indirect taxation on the same tax base follows from the logic of budget constraints and is not specific to this simple model.

[2] For small values of θ, this is equivalent to the constancy of $(\tau-\theta)$.

the real wage paid by firms (w/s) and the real wage received by workers [$(1-\tau)w/(1+\theta)s$] are unaffected, so output is constant. The price level must change, however. Let $L(Y)$ define the demand for money balances. Equilibrium requires that

$$L(Y) = M/s(1+\theta) \tag{7}$$

where M is the nominal money supply. We have followed the standard practice of assuming that the demand for real money balances, deflated by market prices, depends on real output. Since output is unaffected by the tax shift, the after-tax price level $s(1+\theta)$ will not change, and s will fall. Since w/s remains constant, the nominal wage must fall.[3]

Although shifts between direct and indirect taxes are neutral in this model, increases in either are not. Since changes in the total tax burden have real effects in almost any economic model, studies of the effects of changes in specific taxes, such as those surveyed in Ewald Nowotny (1980), do not shed light on the issues considered here.

The hallmark of Keynesian models is that nominal adjustments require time. Changes in the stock of money or shifts between direct and indirect taxation which have no long-run real effects therefore may have important short-run consequences. Sticky nominal wages are the primary rigidity in most Keynesian models.[4] They arise both in textbook Keynesian models and in contracting models such as those of Stanley Fischer (1977) and John Taylor (1980). Customarily, sticky wages are analyzed by adding a description of wage behavior to the classical model, while deleting the requirement that employment equal desired labor supply. Since both explicit and implicit contracts seem to be denominated in terms of pre-tax wages, we assume pre-tax wage rigidity. Since post-tax wages do not need to adjust to tax shifts, rigidities in $(1-\tau)w$ do not imply that shifts between direct and indirect taxation have real effects.

Consider an increase in θ which does not change $(1-\tau)/(1+\theta)$. Given that the nominal wage cannot fall to clear the labor market, employment must fall or prices must rise. In equilibrium, both occur to some extent since a fall in employment lowers output and therefore requires an increase in $s(1+\theta)$ to satisfy (7). Keeping w constant, (1), (2), and (7) imply that the elasticity of the tax-inclusive price with respect to a tax change is

$$\left.\frac{\partial \log[s(1+\theta)]}{\partial \log(1+\theta)}\right|_{w=\bar{w}} = \frac{-L'f'\bar{w}}{Mf''/(1+\theta)-L'f'\bar{w}} > 0. \tag{8}$$

An increase in indirect taxes is similar to a supply shock, a "self-inflicted wound," since prices rise and output falls. This fall in output results from an increase in real wages facing firms, as producer prices decline while nominal wages remain constant.

We have isolated a clear difference in the empirical implications of models with and without nominal rigidities. A natural way of testing for the existence and importance of these rigidities is to examine the response of prices and output to changes in tax structure, controlling for total revenue collections. These tests, while not totally free of ambiguity, are superior to tests of the relationship between money and output for detecting nominal rigidities. Tax structure changes are more likely to be exogenous policy shocks than are changes in the money stock. King and Plosser argue that changes

[3]Alternative approaches might postulate that money demand depends on households' disposable income, $(1-\tau)Y$, or that money balances should be deflated by an index of consumer prices. In the former case, a revenue-neutral shift towards indirect taxation would reduce prices, while in the latter case, it would raise them. In neither case would real output be affected. N. Gregory Mankiw and Summers (1984) present some evidence suggesting the empirical relevance of the case where money demand depends on household expenditure. Regardless of the money demand specification, tax changes will not affect the price level if monetary policy holds nominal *GNP* constant.

[4]An earlier version of this paper also examines price stickiness and real wage resistance and concludes that changes in the tax mix are also likely to have real effects in the presence of these rigidities.

in the money stock may be endogenous. They establish that most of the observed correlation between money and output arises from changes in the money multiplier, not from changes in the stock of base money. The difficulty with tax changes is that they may have incentive and distributional effects which change real magnitudes; we argue below that these effects are unlikely to explain our empirical findings.

II. Methodology

Our aim is to test for the presence of nominal rigidities by studying whether changes in the tax mix affect prices and output. One approach to doing this would be to specify and estimate a structural model including tax variables. Different tax variables would enter different structural equations. Taxes on individual incomes, for example, would not enter equations describing firm behavior. The existence of nominal rigidities could be tested by examining the cross-equation restrictions which insure that changes in the tax mix are neutral. We do not adopt this strategy because of the difficulty of convincingly specifying a complete macroeconomic model.[5] Rather, we gain robustness at the possible expense of efficiency by directly examining the effects of changes in the tax mix using reduced forms. The reduced form of any model with classical properties should imply that tax mix changes do not affect output and prices.

We estimate two systems of equations. The first consists of three reduced-form equations for the logarithms of prices (p_t), nominal after-tax wages ($\tilde{w}_t = w_t(1-\tau_t)$), and output ($y_t$). The explanatory variables are lagged prices, wages, and output, as well as real government deficits (d_t) and the logarithm of the money stock (m_t). We also include three tax variables. The first is $TTOT$, the sum of the direct and indirect tax rates. The second is $TMIX$, the *difference* between the direct and indirect tax rates. Including both $TMIX$ and $TTOT$ is equivalent to including indirect and direct taxes separately. Since we are interested primarily in the effect of switches between direct and indirect taxes holding their sum constant, however, this specification is more natural. The third tax variable, $OTAX$, is the ratio of tax receipts which we classify as neither direct nor indirect taxes to GNP.[6] This system of reduced-form equations can be written as

$$\begin{bmatrix} p_t \\ \tilde{w}_t \\ y_t \end{bmatrix} = \begin{bmatrix} \alpha_p^1(L) & \alpha_w^1(L) & \alpha_y^1(L) \\ \alpha_p^2(L) & \alpha_w^2(L) & \alpha_y^2(L) \\ \alpha_p^3(L) & \alpha_w^3(L) & \alpha_y^3(L) \end{bmatrix} \times \begin{bmatrix} p_{t-1} \\ \tilde{w}_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \alpha_m^1(L) & \cdots & \alpha_x^1(L) \\ \alpha_m^2(L) & \cdots & \alpha_x^2(L) \\ \alpha_m^3(L) & \cdots & \alpha_x^3(L) \end{bmatrix} \times \begin{bmatrix} m_t \\ d_t \\ TTOT_t \\ OTAX_t \\ TMIX_t \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{bmatrix} \tag{9}$$

where the $\alpha^i(L)$'s are second-order lag polynomials. We found that further lagged variables had little explanatory power. Each equation in the system also includes a time trend and seasonal dummy variables.

The inclusion of contemporaneous variables on the right-hand side of the reduced-form equations is a delicate issue. Since we treat $TMIX$ as exogenous, an assumption defended below, its contemporaneous value appears in all the reduced-form equations. Entering contemporaneous monetary and fiscal policy variables as in (9) is strictly appropriate only if they are exogenous. For-

[5] Previous estimates of the role of tax variables in Keynesian Phillips curves and price equations *presume* the presence of nominal rigidities, and thus do not test for their existence.

[6] Since we include the sum of all taxes as well as the deficit, our specification is consistent with both the view that government expenditure is the appropriate measure of fiscal policy, and the view that the deficit is the appropriate measure.

tunately, the results reported below change only negligibly when contemporaneous values of these variables are dropped from the equations.[7]

In addition to fiscal and monetary policy variables, it is desirable to control for shocks to the money demand equation which influence prices and output. If policy is set so as to offset these shocks, it may be appropriate to use nominal *GNP* as a summary variable for the effects of aggregate demand policies. These considerations led Robert Gordon (1983) to pioneer the use of nominal *GNP* in wage and price equations. While this approach captures velocity shocks, it may capture too much: the disadvantage of including nominal *GNP* is that it may not be a predetermined variable.

We estimate a second system of only two equations for nominal after-tax wages and prices which includes current and lagged nominal *GNP* in place of the deficit and the money supply. This system of equations is given by

$$(10)\quad \begin{bmatrix} p_t \\ \tilde{w}_t \end{bmatrix} = \begin{bmatrix} \beta_p^1(L) & \beta_w^1(L) \\ \beta_p^2(L) & \beta_w^2(L) \end{bmatrix} \begin{bmatrix} p_{t-1} \\ \tilde{w}_{t-1} \end{bmatrix} + \begin{bmatrix} \beta_n^1(L) & \cdots & \beta_x^1(L) \\ \beta_n^2(L) & \cdots & \beta_x^2(L) \end{bmatrix} \times \begin{bmatrix} n_t \\ TTOT_t \\ OTAX_t \\ TMIX_t \end{bmatrix} + \begin{bmatrix} \nu_{1t} \\ \nu_{2t} \end{bmatrix}$$

where n_t is the logarithm of nominal *GNP*. In this system, movements in output for a given nominal *GNP* can be calculated from price movements.

Systems (9) and (10) allow for unrestricted wage, price, and output responses to shifts between direct and indirect taxation. Both Keynesian and classical models imply, however, that revenue-neutral tax switches are neutral in the long run. We therefore impose (and test) long-run neutrality by restricting the sum of the *TMIX* coefficients in each equation to equal zero.[8] Imposing long-run neutrality of *TMIX* sharpens the rejection of the hypothesis that *TMIX* is unimportant. The short-run tax neutrality hypothesis implies the restrictions $\alpha_x^1(L) = \alpha_x^2(L) = \alpha_x^3(L) = 0$ in system (9) and $\beta_x^1(L) = \beta_x^2(L) = 0$ in system (10). As long as *TMIX* is a valid exogenous variable, rejection of these null hypotheses is very unfavorable to the classical model. In Section IV we consider some (in our view unlikely) reasons why *TMIX* might appear to matter even if wages and prices were perfectly flexible.

After rejecting these null hypotheses, we focus on the relevance of these rejections for the presence of nominal rigidities. If nominal rigidities are present, then prices should rise and output should fall for some time after a tax switch. To investigate these dynamic effects, we compute our systems' predicted responses to a once and for all decrease in *TMIX*.[9]

The reduced forms described above may be subject to some of the criticisms which have been directed at the vector autoregression approach of Christopher Sims. We have not posited an explicit structural model, and the parameters in our reduced forms might vary with changes in the policy regime. We use our reduced forms, however, only to estimate the effects of certain policy changes

[7] For our purpose—estimating the effects of changes in an exogenous variable—it is of no consequence whether the reduced form is represented as in (9) or triangularized as in Christopher Sims (1980).

[8] This way of imposing long-run neutrality is only valid if prices, wages, and output are stationary while *TMIX* has a unit root. We could not reject either of these hypotheses.

[9] We also followed the spirit of Frederic Mishkin's (1979) approach. This involves fitting a univariate autoregression for *TMIX*. Starting from a base path for *TMIX*, we constructed a new path for *TMIX* by simulating the effect on *TMIX* of a residual in this autoregression. We then compared the paths of prices, wages, and output resulting from these two paths for *TMIX*. This procedure avoids the problems which might arise if permanent shocks to *TMIX* are widely at variance with the historical experience. Because the results of the two approaches were very similar, we report only the permanent shock results.

within a given policy regime. Our view is not that our equations explain how *TMIX* could be used as a major tool of stabilization policy, or even describe the effects of radical changes in *TMIX* outside the sample experience. Rather, we believe that the estimated response of prices and output to changes in *TMIX*, given the current policy regime, can shed light on the existence of nominal rigidities.

Any argument of this type must confront issues similar to those raised in the decades-long debate about the relationship between money and output. The essential identification problem there involves the possibility that money and output are correlated either because they both respond to some third factor, or because changes in money are caused by expectations of changes in output. After presenting our empirical results, we present some evidence supporting the exogeneity of tax changes. At a minimum, it seems clear that changes in the tax mix are much closer to the ideal experiment for studying nominal rigidities than money supply changes.

III. The Data

This section describes our measures of the direct and indirect tax burden in Great Britain and the United States. It begins by discussing conceptual measurement issues which apply to both countries and then considers the data for each nation in some detail.

Direct taxes are defined as taxes on individuals, including income taxes and employee contributions for social insurance. Indirect taxes are those collected from firms. They include sales and value-added taxes, employer contributions for social insurance, and various excise taxes. Our measured tax rates, $\tilde{\tau}$ and $\tilde{\theta}$, are defined as direct and indirect tax receipts as a share of *GNP* at market prices. These variables do not correspond precisely to the proportional tax rates, τ and θ, of Section I. If there were proportional taxes, income tax receipts would equal τY, indirect tax receipts would be $\theta(G+E)$, and Gross National Product measured at market prices would be $(1+\theta)(E+G)$, where G denotes government expenditure. Therefore, our measured tax rates would be

$$(11) \quad \tilde{\tau}=\frac{\tau Y}{(1+\theta)(G+E)}=\frac{\tau}{1+\theta}$$

and

$$(12) \quad \tilde{\theta}=\frac{\theta(G+E)}{(1+\theta)(G+E)}=\frac{\theta}{1+\theta}.$$

Both measured tax rates would be slightly lower than the proportional taxes in our stylized economy. This would bias our measurement of *TMIX*, since

$$(13) \quad TMIX=\tilde{\tau}-\tilde{\theta}=(\tau-\theta)/(1+\theta).$$

For values of θ between 0 and .15, however, as in our sample, this bias would be small. In contrast, the measured tax rates would yield exactly the correct measure for the total tax burden, *TTOT*:

$$(14) \quad TTOT=\tilde{\tau}+\tilde{\theta}=(\tau+\theta)/(1+\theta).$$

In Section I we analyzed tax reforms which altered τ or θ while keeping $(1-\tau)/(1+\theta)$ constant. Since $1-(1-\tau)/(1+\theta)=(\tau+\theta)/(1+\theta)$, a tax reform with no effect on $(1-\tau)/(1+\theta)$ will not change *TTOT*.

Our approach to measuring tax rates is only one of many possibilities. Ideally, we would like our tax variables to be legislated tax rates which change only when government policy changes. Unfortunately, taxes are too complex for us to define either *the* direct tax rate or *the* indirect tax rate. The tax base is much smaller than *GNP* and taxes are frequently raised or lowered by changing the tax base. Direct taxes are not strictly proportional to income, nor indirect taxes proportional to expenditure. In particular, indirect taxes do not cover all goods.

If the elasticities of direct and indirect tax receipts with respect to *GNP* are different, then measured *TTOT* and *TMIX* variables may be affected by cyclical fluctuations. This could induce a spurious correlation between the tax variables, prices, and output. We therefore employ another technique for identifying shifts between direct and indirect

taxation. We construct measures of full-employment *TMIX* and *TTOT* using data on full-employment tax receipts and *GNP* to avoid problems due to cyclical fluctuations.

A. *The United Kingdom*

Direct taxes in the United Kingdom consist of personal income taxes and surtaxes and employees' national insurance contributions. Indirect taxes include a variety of different levies: Purchase Tax (prior to 1972), Value-Added Tax, stamp, customs, alcohol, and tobacco duties, car tax, as well as employers' contributions for National Insurance and Selective Employment Tax. Data on tax receipts were obtained from *Financial Statistics* and the Central Statistical Office. A detailed data description is available from the authors on request.

The resulting shares of direct and indirect taxes in *GDP* are plotted in Figure 1. The share of indirect taxes ranges from just over 11 percent in 1963 to more than 15 percent during the early 1980's. There are even more significant movements in the direct tax share, which varies between 10.1 and 16.8 percent. The figure also shows that there are some tax reforms which correspond to shifts between the two sources of revenue. In particular, the 1979 tax reform involved a reduction of basic statutory income tax rates accompanied by systematic increases in *VAT*. The direct tax cuts were forecast to reduce revenue by 4.5 billion pounds, while the increase in *VAT* was expected to raise 4.2 billion pounds. This is the cleanest example of a tax reform which changed the "side of the market" on which taxes are levied.

We measure the British price level using the deflator for *GDP* at market prices.[10] Our nominal wage measure is the index of basic weekly wage rates in all industries and services. Output is measured by real *GDP* at market prices. Our equations also include the logarithm of *M*1, the deficit as measured by the Public Sector Borrowing Requirement

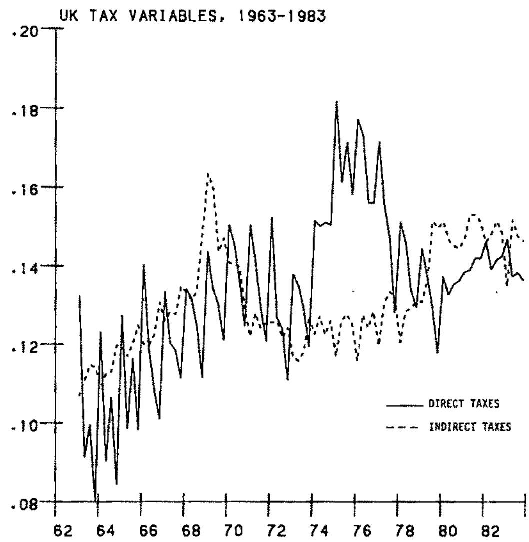


FIGURE 1. U.K. DIRECT AND INDIRECT TAXES AS A FRACTION OF *GDP*

(*PSBR*), and the level of other tax receipts, defined as total government tax receipts less direct and indirect taxes, divided by *GDP*.[11]

There are several intervals of statutory wage and price controls and implicit wage-price guidelines during our sample period. Previous attempts to find significant effects from price controls (for example, J. D. Sargan, 1980) have been unsuccessful. Wage controls do appear to have affected wage growth, however. S. G. B. Henry (1984) identifies five periods of statutory wage restraint and associated wage catch-up. We include his set of indicator variables for wage controls in all of our reduced-form equations.[12]

[10] Using the Retail Price Index to measure prices did not change our results.

[11] Quarterly *PSBR* and *M*1 data are only available since 1963, and the wage series that we use was not computed after 1983. Our sample period is therefore limited to the 84 quarters between 1963:1 and 1983:4.

[12] There is some disagreement regarding the most binding periods of wage control. Gordon uses dummy variables which differ from those in Henry, and Sushil Wadhwani (1983) uses yet another set. Our results were insensitive to alternative choices. We amend Henry's variables by adding an indicator variable for rapid wage growth in the second quarter of 1978.

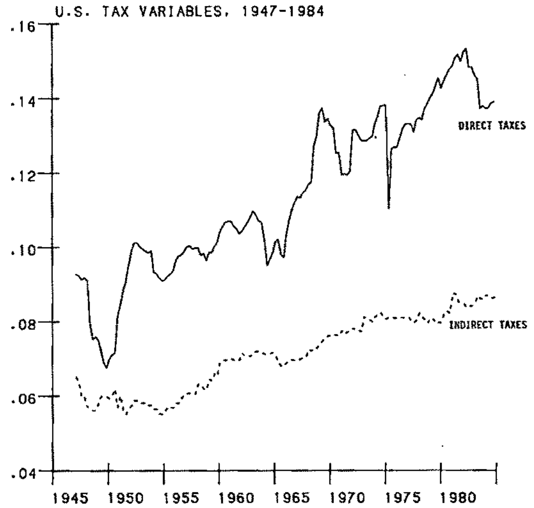


FIGURE 2. U.S. DIRECT AND INDIRECT TAXES AS A FRACTION OF *GNP*

B. *The United States*

Direct tax receipts for the United States include federal personal income tax receipts, state and local personal income tax receipts, and personal contributions for social insurance. Our measure of indirect taxes is the sum of federal indirect business taxes, which consist of both excise taxes and customs duties, state and local sales tax receipts, and private employer contributions for social insurance.[13] Direct and indirect tax receipts as a fraction of *GNP* are shown in Figure 2. The share of direct taxes in *GNP* displays substantial variability in the postwar period, ranging from only 7 percent in 1949 to nearly 15 percent early in the 1980's. Indirect taxes are much less volatile, ranging between 5.7 and 8.7 percent of *GNP* and trending upward throughout the sample period.

We measure the U.S. price level using the *GNP* deflator.[14] Wages are measured as average hourly earnings in manufacturing, and output as *GNP* in 1972 dollars. Our equations include the logarithm of *M*1, the level of the total government deficit, and other tax receipts, defined as total tax receipts less direct and indirect taxes, divided by *GNP*. We also include two variables drawn from Gordon and Stephen King (1982) to allow for the impact of wage and price controls during the early 1970's.

IV. Empirical Findings

This section reports estimates of how switches between direct and indirect taxation affect wages, prices, and output. We consider the British and the American experience in turn. The section closes with several qualifications to our findings.

A. *The United Kingdom*

Equation systems (9) and (10) are estimated using data for the 1963–83 period. The coefficient estimates are reported in Table 1. Both systems suggest that changes in the direct vs. indirect tax mix have substantial effects. The null hypothesis that the *TMIX* coefficients equal zero is rejected decisively in each case. The test statistic in system (9) is 32.5; it is distributed $\chi^2(6)$ under the null hypothesis that the tax mix variables have no effect on the short-run movements in wages, prices, and output. The null hypothesis is rejected at the .01 level. For system (10), the test statistic is 27.6. In this case, with only two equations, the test statistic is distributed $\chi^2(4)$ under the null hypothesis; again, we reject the neutrality hypothesis at the .01 level. These overwhelming rejections suggest the potential importance of nominal rigidities. On the other hand, the *long-run* neutrality of *TMIX* is not rejected at conventional significance levels. By comparing the likelihood of (9) to that of a system that does not impose the restriction that the sum of the *TMIX* coefficients in each equation equal zero, we obtain a test statistic of 3.7 which is distributed as $\chi^2(3)$ under the hypothesis of long-run neutrality.

To describe the effect of raising indirect taxes, we compute impulse response func-

[13] We exclude state and local government employer contributions from our calculation of social insurance contributions by employers.

[14] Our results were unchanged when we used the *CPI*-Urban Worker price index, excluding shelter, to measure the price level.

TABLE 1—REDUCED-FORM COEFFICIENTS ESTIMATES, UNITED KINGDOM

Independent Variable	Equation System (9) Price	Equation System (9) Nominal After-Tax Wage	Equation System (9) Real *GDP*	Equation System (10) Price	Equation System (10) Nominal After-Tax Wage
Constant	−1.1730	0.568	2.993	−2.050	0.432
	(0.990)	(1.165)	(1.1724)	(0.910)	(1.137)
$Price_{-1}$	1.122	0.434	−0.489	0.955	0.501
	(0.149)	(0.176)	(0.260)	(0.133)	(0.166)
$Price_{-2}$	−0.261	−0.485	−0.055	−0.163	−0.503
	(0.132)	(0.155)	(0.230)	(0.108)	(0.135)
$Wage_{-1}$	0.189	1.213	0.048	0.150	1.236
	(0.094)	(0.111)	(0.164)	(0.085)	(0.106)
$Wage_{-2}$	−0.043	−0.249	0.319	−0.111	−0.236
	(0.092)	(0.109)	(0.161)	(0.085)	(0.107)
$Real\ GDP_{-1}$	0.095	−0.048	0.128	–	–
	(0.077)	(0.091)	(0.134)		
$Real\ GDP_{-2}$	−0.0003	−0.017	0.323	–	–
	(0.087)	(0.103)	(0.152)		
$TMIX - TMIX_{-1}$	−0.434	−0.639	0.037	−0.304	−0.672
	(0.166)	(0.195)	(0.289)	(0.143)	(0.179)
$TMIX - TMIX_{-2}$	0.121	0.121	−0.042	0.114	0.186
	(0.149)	(0.175)	(0.259)	(0.131)	(0.163)
$TTOT$	0.057	0.150	−0.512	0.179	1.233
	(0.175)	(0.206)	(0.305)	(0.155)	(0.191)
$TTOT_{-1}$	−0.141	−0.552	0.468	−0.124	−0.588
	(0.163)	(0.193)	(0.294)	(0.146)	(0.183)
$TTOT_{-2}$	0.049	0.329	0.143	−0.020	0.175
	(0.182)	(0.214)	(0.316)	(0.153)	(0.193)
$OTAX$	−0.047	−0.004	0.124	−0.098	−0.033
	(0.083)	(0.098)	(0.145)	(0.073)	(0.092)
$OTAX_{-1}$	0.063	−0.030	0.373	−0.074	−0.036
	(0.086)	(0.102)	(0.152)	(0.075)	(0.093)
$OTAX_{-2}$	−0.068	−0.007	−0.048	−0.043	−0.003
	(0.082)	(0.096)	(0.142)	(0.069)	(0.086)
$DEF(\times 10^{-6})$	1.950	0.148	1.502	–	–
	(1.580)	(1.860)	(2.751)		
$DEF_{-1}(\times 10^{-6})$	−1.256	1.983	−0.088	–	–
	(1.640)	(1.930)	(2.856)		
$DEF_{-2}(\times 10^{-6})$	0.566	1.786	0.164	–	–
	(1.541)	(1.814)	(2.684)		
$MONEY$	−0.086	−0.023	0.083	–	–
	(0.072)	(0.085)	(0.125)		
$MONEY_{-1}$	−0.031	−0.034	−0.060	–	–
	(0.098)	(0.116)	(0.171)		
$MONEY_{-2}$	0.091	0.083	0.099	–	–
	(0.070)	(0.082)	(0.121)		
$NOMGDP$	–	–	–	0.200	−0.041
				(0.072)	(0.080)
$NOMGDP_{-1}$	–	–	–	0.041	−0.004
				(0.072)	(0.089)
$NOMGDP_{-2}$	–	–	–	−0.059	−0.014
				(0.081)	(0.101)
SSR	.0037	.0052	.0113	.0036	.0056
$Q(4)$	7.095	4.738	8.430	4.512	3.355
SEE	.0089	.0105	.0155	.0100	.0104

Notes: All equations are estimated using quarterly data for the 1963:3–1983:4 period, for a total of 82 observations. Standard errors are shown in parentheses. Each equation also includes a time trend, seasonal dummy variables, and wage and price control dummy variables. See the text for further description.

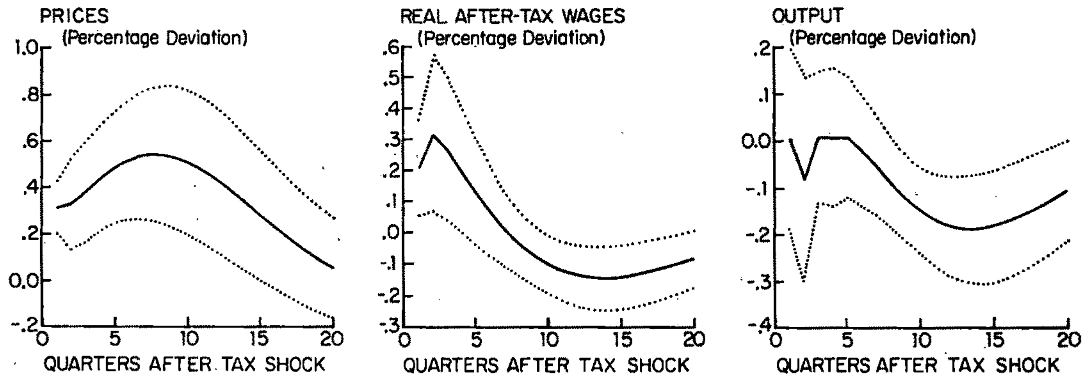


FIGURE 3. IMPULSE RESPONSE FUNCTIONS FOR PERCENTAGE DEVIATION IN EACH VARIABLE (Equation System UK–9)

tions for prices, real after-tax wages, and output with respect to a 1 percent decrease in *TMIX*. This corresponds to an indirect tax increase of one-half of 1 percent of *GDP*, accompanied by an equal-revenue reduction in direct taxes. If all pre-tax prices remained fixed, the tax-inclusive price level would rise by one-half of 1 percent. The impulse response functions for system (9), labelled UK–9, are shown in Figure 3. The figure reports both the point estimates of the percentage deviation of each variable from its initial steady-state value, as well as one standard error bands.[15]

In the quarter when the tax change occurs, prices rise by three-tenths of 1 percent. They continue to rise for eight quarters thereafter, peaking .54 percent above their initial level. Prices then decline, but remain more than .1 percent above their initial value for four and one-half years after the tax change. For the first five quarters after the shock, the sum of the deviations of the price level from its initial value is 2.04, with a standard error of 0.97. The null hypothesis of no price effects over this horizon is rejected at the .05 level. Similarly, over a ten-quarter horizon, the sum of the price effects is 4.56, with a standard error of 2.24.

Nominal after-tax wages also rise after the tax change. In the first quarter, they increase by nearly half a percent, raising the firm's real wage by .2 percent. The real wage increases for another quarter and then begins to decline. By seven quarters after the tax reform, real wages have fallen below their initial level and they remain more than .1 percent below their starting point for nearly two years. The figure shows that the wage dynamics are not as well determined as those for prices. The sum of the wage impulses for the first five quarters after the shock is 0.99, with a standard error of 0.90.

The impulse response path also shows output moving erratically. The estimates of the output response function are imprecise, however. Output rises in the quarter when the shock occurs, and then declines in the next quarter. The sum of output deviations for the five quarters after the shock is -0.07 percent, with a standard error of .61. By ten quarters after the shock, the comparable value is $-.310$ with a standard error of .465. Six quarters after the tax shock, output enters a long period of decline. At the lowest point on its trajectory, output is .19 percent below its initial level. The individual quarter output effects should be regarded with caution, however, as the large standard errors suggest.

[15] The impulse response function standard errors are computed using standard asymptotic methods. Defining $f(\hat{\alpha}, t)$ as the impulse response function t periods after the shock, and $\hat{\alpha}$ the coefficient estimates from (9), we compute the variance of $f(\hat{\alpha}, t)$ as $\nabla f' \, \Omega \nabla f$, where ∇f is the vector of derivatives of f with respect to $\hat{\alpha}$, and Ω is the covariance matrix of $\hat{\alpha}$.

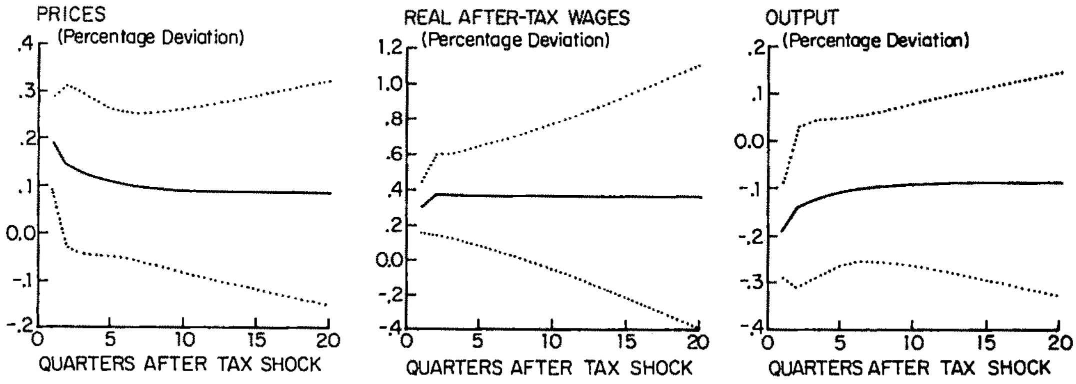


FIGURE 4. IMPULSE RESPONSE FUNCTIONS FOR PERCENTAGE DEVIATION IN EACH VARIABLE (Equation System UK–10)

The estimates from system (10) also suggest significant tax effects, as can be seen from the impulse response functions in Figure 4, labelled UK–10. Prices rise by .19 percent in the quarter of the shock, 40 percent of the amount which would be predicted if pre-tax prices were completely fixed. They decline slowly thereafter, and are still more than .08 percent above their initial level four years afterwards. The standard errors for the impulse response functions from (10) are, however, larger than those from (9). Five quarters after the tax change, the sum of the deviation of prices from their initial level is .66 percent, with a standard error of .71. Real wages again rise for a short while after the tax shock occurs, then decline. Output changes in this system, which are equal to the negative of the price impulses, display a more stable response pattern than those in system (9).

We explored the robustness of our tax mix results in several ways. We added exchange rates as additional explanatory variables; they had statistically insignificant coefficients and did not affect our rejection of short-run tax neutrality. We also estimated our equations without the indicator variables for wage and price controls, and most of the tax coefficients changed very little. Adding further lagged variables to the system reduced the statistical significance of some coefficient estimates, but had little impact on either our estimated dynamic responses or our rejections of the tax neutrality hypothesis.[16]

A possible problem with our results is that movements in *TMIX* may reflect changes in economic conditions. In order to explore this possibility, we examined the effects of two specific episodes of tax reform in Britain where the relative importance of direct and indirect taxes was altered. In 1979, the *VAT* was increased and marginal personal income tax rates were cut. In April 1976, value-added taxes on durables were reduced. Replacing *TMIX* by dummy variables for these reforms led to conclusions very similar to those reported in this section.[17]

B. *The United States*

In this section, we investigate whether our U.K. findings are consistent with the U.S. experience from 1948:1 to 1984:3. Estimates

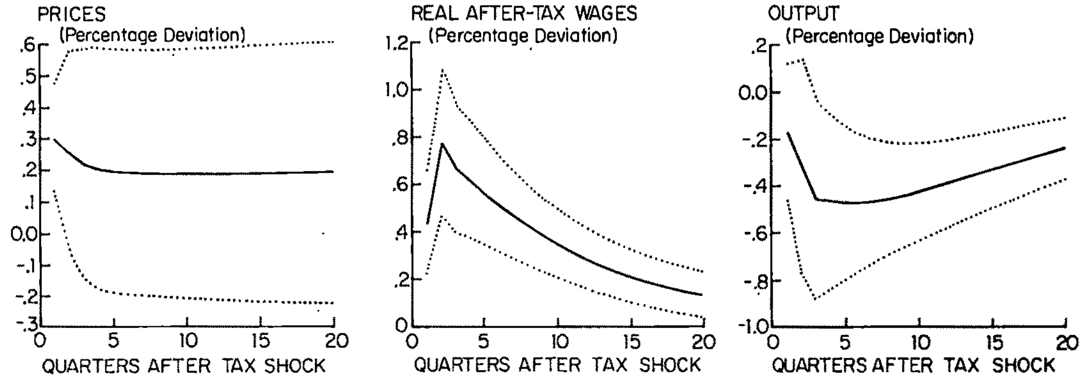
[16]A change in the total tax burden also has real effects. In the three-equation system, a 1 percent of *GDP* increase in the total tax burden reduces output .51 percent in the quarter of the tax change and induces lower output for 3 quarters after the shock. The estimates of *TTOT*'s impact on both prices and output, however, are plagued by very large standard errors.

[17]Details results on the effect of the 1979 tax switch are presented in an earlier version of this paper, available from the authors on request. The inflationary impact of the 1979 tax reform is also discussed in Willem Buiter and Marcus Miller (1981).

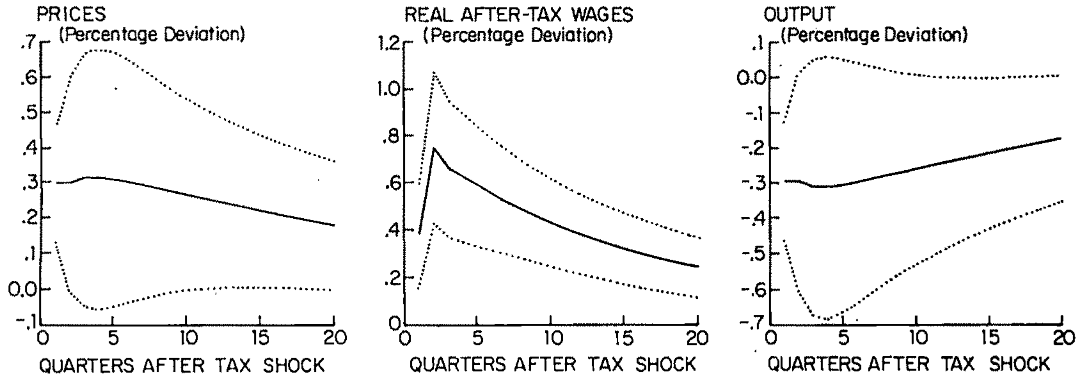
TABLE 2—REDUCED-FORM COEFFICIENT ESTIMATES, UNITED STATES

Independent Variable	Equation System (9): Price	Equation System (9): Nominal After-Tax Wage	Equation System (9): Real *GNP*	Equation System (10): Price	Equation System (10): Nominal After-Tax Wage
Constant	−0.166	0.013	0.691	−0.325	−0.224
	(0.109)	(0.121)	(0.189)	(0.093)	(0.103)
$Price_{-1}$	1.351	0.486	0.149	1.341	0.425
	(0.090)	(0.100)	(0.156)	(0.084)	(0.093)
$Price_{-2}$	−0.349	−0.352	−0.045	−0.394	−0.393
	(0.094)	(0.104)	(0.162)	(0.082)	(0.091)
$Wage_{-1}$	−0.040	0.719	−0.345	0.056	0.874
	(0.086)	(0.095)	(0.148)	(0.080)	(0.089)
$Wage_{-2}$	0.032	0.169	0.235	−0.041	0.066
	(0.080)	(0.089)	(0.139)	(0.077)	(0.085)
$Real\ GNP_{-1}$	0.026	0.008	1.005	–	–
	(0.055)	(0.061)	(0.095)		
$Real\ GNP_{-2}$	−0.031	0.002	−0.113	–	–
	(0.051)	(0.057)	(0.089)		
$TMIX - TMIX_{-1}$	−0.421	−0.380	0.235	−0.434	−0.329
	(0.231)	(0.255)	(0.399)	(0.218)	(0.241)
$TMIX - TMIX_{-2}$	0.119	−0.358	−0.064	0.137	−0.338
	(0.194)	(0.215)	(0.336)	(0.180)	(0.198)
$TTOT$	0.377	0.641	−0.818	0.423	0.758
	(0.186)	(0.206)	(0.321)	(0.175)	(0.193)
$TTOT_{-1}$	−0.351	−0.277	0.766	−0.478	−0.426
	(0.252)	(0.280)	(0.437)	(0.228)	(0.252)
$TTOT_{-2}$	0.063	−0.271	−0.115	0.153	−0.219
	(0.192)	(0.212)	(0.332)	(0.172)	(0.190)
$OTAX$	0.694	0.829	0.952	0.302	0.459
	(0.151)	(0.167)	(0.261)	(0.167)	(0.185)
$OTAX_{-1}$	−0.301	−0.098	−0.031	−0.295	−0.167
	(0.209)	(0.231)	(0.362)	(0.196)	(0.217)
$OTAX_{-2}$	−0.152	−0.085	−0.048	−0.057	0.064
	(0.157)	(0.173)	(0.271)	(0.147)	(0.162)
$DEF(\times 10^{-5})$	4.295	1.180	3.197	–	–
	(4.559)	(5.049)	(7.892)		
$DEF_{-1}(\times 10^{-5})$	−2.888	−7.512	−24.096	–	–
	(7.371)	(8.163)	(12.761)		
$DEF_{-2}(\times 10^{-5})$	7.188	5.429	−0.183	–	–
	(4.983)	(5.518)	(8.625)		
$MONEY$	0.107	0.146	0.298	–	–
	(0.052)	(0.057)	(0.089)		
$MONEY_{-1}$	−0.160	−0.227	−0.190	–	–
	(0.082)	(0.091)	(0.143)		
$MONEY_{-2}$	0.090	0.077	−0.102	–	–
	(0.056)	(0.062)	(0.092)		
$NOMGNP$	–	–	–	0.207	0.232
				(0.041)	(0.046)
$NOMGNP_{-1}$	–	–	–	−0.200	−0.259
				(0.066)	(0.073)
$NOMGNP_{-2}$	–	–	–	0.044	0.069
				(0.045)	(0.050)
SSR	.0022	.0027	.0066	.0022	.0027
$Q(4)$	7.855	1.348	6.713	8.029	3.334
SEE	.0043	.0047	.0074	.0042	.0046

Notes: All equations are estimated using quarterly data for the period 1948:1–1984:3, a total of 147 observations. Standard errors are shown in parentheses. Each equation also includes a time trend, seasonal dummy variables, and wage and price control dummy variables. See the text for further description.



Panel A: Equation System US–9



Panel B: Equation System US–10

FIGURE 5. IMPULSE RESPONSE FUNCTIONS FOR PERCENTAGE DEVIATION IN EACH VARIABLE

of both sets of equations are reported in Table 2. The central question is whether we can reject the null hypothesis that *TMIX* should be excluded from systems (9) and (10). For system (9), the test statistic is 21.7. Since it is distributed $\chi^2(6)$ under the short-run tax neutrality hypothesis, this constitutes a rejection at the .01 level. For system (10), the two-equation system, the test statistic of 17.8, which is distributed as $\chi^2(4)$ under the null, also implies rejection at the .01 level. These findings provide strong evidence for the presence of wage or price stickiness in the United States.

To test long-run neutrality, we again compare the likelihood under the assumption that the sum of the coefficients of *TMIX* is zero to that obtained when relaxing this condition. Here we obtain a test statistic of 8.6, which is distributed $\chi^2(3)$ under the hypothesis of long-run neutrality. Thus this hypothesis is rejected at the 5 percent level, but not the 1 percent level.

The two panels of Figure 5, respectively labelled US–9 and US–10, report the *TMIX* impulse response functions corresponding to systems (9) and (10) for the United States. The initial effect of a permanent 1 percent *TMIX* decrease is a .32 percent increase in prices, just over 60 percent of the impact with fixed producer prices. The absence of significant tax variation makes the standard errors on the estimated price responses larger than those for Britain. The sum of the price changes for the first five quarters after the change is 1.13, with a standard error of 1.63.

The American evidence also differs from the British in suggesting much slower adjustment back to equilibrium.

Real wages also rise after a tax shock, corroborating our British findings. The initial effect of a 1 percent *TMIX* drop is to raise the firm's real wage by .44 percent. Real wages continue to increase for one additional quarter and then decline monotonically to their initial level. Adjustment is slow; even five years after the shock, real wages are .13 percent above their initial level.

Output experiences a pronounced decline after an increase in indirect taxation. A 1 percent fall in *TMIX* induces a .2 percent drop in real *GNP* in the quarter of the tax change. The path of output thereafter depends upon the choice between systems (9) and (10). In (9), output continues to decline for another quarter and falls to .45 percent below its initial level before starting to return to its initial level. The sum of the output effects up to ten quarters after the change is -4.12, with a standard error of 2.73. The results for (10) suggest that the amount of lost output declines after the first quarter, although output returns to its initial level very slowly. The ten-quarter sum equals -2.90 (2.95). Both sets of results are consistent with the view that nominal wages are sticky, since the insufficient nominal wage decline in response to indirect tax increases raises real wages and induces firms to lay off workers. This has the ultimate effect of lowering real money balances.

Our findings are insensitive to several specification changes. Excluding Gordon and King's wage-price control variables has little effect on the estimated coefficients and impulse response functions. Adding interest rates, exchange rates, and further lagged values of the currently included variables also has little substantive impact on our conclusions. The central finding, that the short-run tax neutrality hypothesis is strongly rejected, obtains in a wide variety of specifications.

These results can also be used to study the impact of revenue-raising tax increases. A permanent increase in the total tax burden, keeping *TMIX* constant, increases prices and real wages and causes a drop in output. A 1 percent increase in *TTOT* raises prices by .38 percent and real wages by .26 percent in the first quarter. Output declines by .81 percent when the shock occurs and continues to fall thereafter. By eight quarters after the tax increase, output is 1.65 percent below its starting value. These findings, while suggestive, are accompanied by large standard errors and should therefore be interpreted with caution.

C. *Qualifications*

Two potentially important assumptions underlie our use of the *TMIX* variable to test for the existence of nominal rigidities. First, we assume that *TMIX* is exogenous in our reduced-form systems. Second, we postulate that except for the effects of wage and price stickiness, changes in *TMIX* should have no impact on prices or output. The possible failure of parallel assumptions has caused debate about the interpretation of linkages between money and output. We consider each assumption in turn.

Several arguments for the endogeneity of our tax mix variable might be constructed. Perhaps most plausibly, it might be noted that if the output elasticities of direct and indirect taxes are different, then changes in real output will induce changes in *TMIX*. Price shocks may be transmitted to *GNP* and then to *TMIX* as well. This issue is partly addressed by our inclusion of lagged output in the reduced-form system, and by our separate examination of the 1979 and 1976 policy changes in Great Britain. As a further check, we use data on cyclically adjusted revenue collections to create full-employment *TMIX* and *TTOT* variables for the United States.[18] These data were only available for the post-1955 period. The results obtained using these variables were similar to those obtained with our unadjusted tax variables, suggesting that cyclical fluctuations are not an important source of

[18]Full-employment data are not available for the U.K. on a quarterly basis. In the United States, data on federal taxes beginning in 1955 are published in Thomas Halloway (1984). Estimates of high-employment state and local receipts were constructed by the authors.

endogeneity for the receipts-based tax measures.[19] Unfortunately, the data are not available to examine the effects of cyclical adjustments for Great Britain, or for the entire post-1948 period in the United States.

An alternative argument against the exogeneity of *TMIX* might hold that the tax mix is set in response to projected economic conditions, or that it helps to forecast future economic policies. The historical context which generated changes in *TMIX* does not support these views. The 1979 tax reform in Great Britain immediately followed an election which was decided on grounds other than tax policy. The avowed purpose of its proponents was to improve incentives through reductions in marginal income tax rates. In the United States, most of the variation in indirect taxes comes from movements in state sales taxes and employer payroll taxes. Neither of these are likely to be manipulated for macroeconomic purposes. More generally, it seems unlikely that governments systematically shift towards indirect taxes when they foresee rising prices or when they intend to pursue more expansionary monetary policy. The 1979 reform in Britain was accompanied by an announced policy of monetary restraint. Nothing in the history of either British or American tax policy suggests that tax changes should help to forecast future monetary policies. This inference is consistent with the failure of Granger causality tests to reject the hypothesis that *TMIX* does not cause either money or *TTOT*.

A second potential objection to our tests is that *TMIX* might have effects on output and prices through channels other than wage and price rigidities. Such a possibility cannot be ruled out, since changes in *TMIX* do not correspond precisely to our theoretical model. Indirect taxes do not cover all goods, and direct taxes are not strictly proportional. Nonetheless, it is difficult to explain our findings along these lines. Increases in indirect taxes coupled with equal revenue decreases in direct taxes are usually thought to improve incentives to work and invest. Since indirect taxes are also less progressive than direct taxes, they should have smaller disincentive effects. Thus they should raise output and reduce prices—the opposite of what we find.

There are no controlled experiments in macroeconomics. Nevertheless, we find it difficult to account for our results in terms of the limitations of tax-shift experiments. At a minimum, the flaws in our tax-based tests are largely independent of those in tests which focus on the relationship between money and output. Our tests therefore provide at least some additional evidence to support the hypothesis of wage and price stickiness.

V. Conclusions

A major thrust of much recent macroeconomic research has been the elucidation of business cycles as equilibria of competitive economies with fully flexible prices. Theories in both the "misperceptions" and "real business cycle" traditions emphasize the assumption of perfect price flexibility and the resulting absence of unexploited opportunities for beneficial exchange. These theories imply strong data restrictions: fully perceived changes in government policy which do not change any agent's opportunity set should have no real effects. In contrast, the essence of contemporary Keynesian thinking is that prices are in some sense sticky, so some purely nominal disturbances do matter.

The difficulty in empirically distinguishing these theories arises from the problem of isolating purely nominal disturbances. Traditionally, they have been tested by examining the relationship between variously measured monetary shocks and real variables. These tests have not been entirely conclusive because a variety of rationalizations, with very different structural implications, can be offered for the comovement of money and output.

[19]Although the results using full-employment and unadjusted *TMIX* are always similar, the resemblance between our equations for the 1948–84 period and the comparison equations for 1955–84 depended upon our choice of price series. The equations using the shelter-exclusive *CPI* are very similar to those for the full-sample period, while those using the *GNP* deflator are substantially different.

In this paper, we rely on tax shocks of a special sort to distinguish between classical and Keynesian models. A clear implication of microeconomic theory with flexible prices is that the side of the market on which a tax is collected does not influence its ultimate real effects. Tax changes between direct and indirect taxation therefore provide a natural experiment for examining the importance of nominal rigidities. The appeal of the experiment is enhanced by the apparently unsystematic way in which taxes have varied.

The results of our investigation provide evidence against the classical view that wages and prices are perfectly flexible. While arguments may be made to rationalize the comovements we observe with perfectly flexible prices, we find it impossible to convincingly account for the empirical regularities in the data without assuming some sort of price rigidity.

Asserting that prices are rigid falls far short of explaining them or understanding their properties. Our results suggest that this remains a vitally important research problem. "Menu costs," which have been proposed as one explanation for price rigidities, cannot explain why many prices which can be changed at low cost, such as newsstand magazine prices,[20] appear to change infrequently. Moreover, monetary policy appears potent even in highly inflationary economies, where menu costs should be less important.

Our results have potentially important consequences for tax policy. Almost universally, reforms in the tax structure are evaluated within the context of market-clearing models where prices are perfectly flexible. Within such models, the distinction between direct and indirect taxation is of no consequence. Our findings suggest that this distinction may be important over periods of several years, during which prices are not fully flexible. Indeed, the macroeconomic consequences of some reforms may dwarf their microeconomic impact on economic efficiency. If unemployment is a significant byproduct of certain tax reforms, traditional thinking about their incidence needs to be reconsidered.

Consider as an example current proposals to raise revenue by taxing domestic and imported crude oil. Estimates by the Congressional Budget Office (1985) suggest that this measure would raise about $4.2 billion for each $1 per barrel tax. Thus a $5 a barrel tax would raise the indirect tax burden by $21 billion. Our estimates suggest that if monetary policy were not altered, this would result in lost output of $60 billion over the succeeding decade. Similar estimates are obtained assuming that monetary policy acts to keep nominal *GNP* constant following the tax reform. These figures bulk large relative to the allocative effects traditionally emphasized in microeconomic analyses of excise tax reforms.

Some might argue that it is inappropriate to assess the output effects of tax reforms while holding monetary policy constant, since monetary policy could accommodate tax changes. This issue is treated in our paper (in preparation). If the monetary authority has set monetary policy to trade off unemployment and inflation in a desirable way prior to tax reform, however, the loss of welfare from a small tax change will be independent of the monetary policy response. Unless one believes that monetary policy is wrong prior to a tax reform, there is no reason not to evaluate the effects of the tax holding monetary policy constant. This is especially true for small reforms in excise taxation. The effects of these reforms would be difficult to disentangle accurately enough for them to be explicitly accommodated by monetary policy.

Our finding that shifts towards indirect taxation have adverse macroeconomic consequences raises an obvious question. Could macroeconomic performance be improved by reducing indirect taxes and increasing direct taxes? The conscious and regular use of such tax policies as stabilization measures would be such a significant change in policy regime that our estimates cannot shed much light on this issue. They do suggest, however, that such a change might well be desirable on a one-shot basis. The gains might be taken

[20] Stephen Cecchetti (1984) presents detailed evidence on the inflexibility of magazine prices.

either in the form of reduced inflation or increased output. Our paper (in preparation) demonstrates that if output is held constant, tax changes may have a permanent effect on the rate of inflation.

Our results in this paper suggest a number of directions for future research. The robustness of our conclusions might be examined by studying tax changes in other countries or in individual American states. Structural estimation might yield more precise information on the nature of wage and price stickiness, and tax reforms might facilitate identification of these models. The effects of alternative policy responses to large tax reforms might also be considered. Perhaps most importantly, our results isolate a major class of apparent rigidities which economic theory needs to explain.

REFERENCES

Break, George M., "The Incidence and Economic Effects of Taxation," in Alan Blinder et al., eds., *The Economics of Public Finance*, Washington: The Brookings Institution, 1974.

Buiter, Willem and Miller, Marcus, "The Thatcher Experiment: The First Two Years," *Brookings Papers on Economic Activity*, 2:1981, 315–79.

Cecchetti, Stephen, "The Frequency of Price Adjustment: A Study of the Newsstand Prices of Magazines," mimeo., NYU Graduate School of Business, 1984.

Fischer, Stanley, "Long Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *Journal of Political Economy*, February 1977, *86*, 191–206.

Gordon, Robert J., "A Century of Evidence on Wage and Price Stickiness in the United States, the United Kingdom, and Japan," in J. Tobin, ed., *Macroeconomics, Prices and Quantities*, Washington: The Brookings Institution, 1983.

______ **and King, Stephen,** "The Output Cost of Disinflation in Traditional and Vector Autoregressive Models," *Brookings Papers on Economic Activity*, 1:1982, 205–42.

Halloway, Thomas M., "Cyclical Adjustment of the Federal Budget and Federal Debt: Detailed Methodology and Estimates," U.S. Department of Commerce, Bureau of Economic Analysis Staff Paper 40, Washington, 1984.

Henry, S. G. B., "Incomes Policy and Aggregate Pay," in R. Elliot and L. Fallick, eds., *Incomes Policies, Inflation, and Relative Pay*, London: George Allen and Unwin, 1984.

King, Robert G. and Plosser, Charles, "Money, Credit, and Prices in a Real Business Cycle Model," *American Economic Review*, June 1984, *74*, 363–80.

Mankiw, N. Gregory and Summers, Lawrence, "Are Tax Cuts Really Expansionary?," NBER Working Paper No. 1443, 1984.

Mishkin, Frederic S., "Simulation Methodology in Macroeconomics: An Innovation Technique," *Journal of Political Economy*, August 1979, *87*, 816–36.

Nowotny, Ewald, "Inflation and Taxation," *Journal of Economic Literature*, September 1980, *43*, 1025–50.

Poterba, James, Rotemberg, Julio and Summers, Lawrence, "Tax Policy and Inflation," in preparation.

Sargan, J. D., "The Consumer Price Equation in the Post-War British Economy: An Exercise in Equation Specification Testing," *Review of Economic Studies*, January 1980, *47*, 113–35.

Sims, Christopher A., "Macroeconomics and Reality," *Econometrica*, January 1980, *48*, 1–48.

Solow, Robert M., "On Theories of Unemployment," *American Economic Review*, March 1980, *70*, 1–10.

Taylor, John, "Aggregate Dynamics and Staggered Contracts," *Journal of Political Economy*, February 1980, *88*, 1–23.

Wadhwani, Sushil, "Wage Inflation in the United Kingdom," Working Paper No. 132, Centre for Labour Economics, London School of Economics, 1983.

U.S. Congressional Budget Office, *Reducing the Deficit: Spending and Revenue Options*, Washington: USGPO, 1985.

Ricardian Consumers with Keynesian Propensities

By ROBERT B. BARSKY, N. GREGORY MANKIW, AND STEPHEN P. ZELDES*

This paper examines Ricardian equivalence in a world in which taxes are not lump sum, but are levied on risky labor income. It shows that the marginal propensity to consume out of a tax cut, coupled with a future income tax increase, can be substantial under plausible assumptions. Indeed, the MPC out of a tax cut can be closer to the Keynesian value that ignores the future tax liabilities than to the Ricardian value that treats future taxes as if they were lump sum.

In conventional Keynesian macroeconomic models, a debt-financed tax cut stimulates aggregate demand. An alternative view, first noted by David Ricardo and revived by Robert Barro (1974), is that a tax cut merely replaces current taxes with future taxes of equal present value. If taxes are lump sum, capital markets are perfect, and all individuals have operative altruistic bequest motives, debt and tax finance are equivalent, and tax cuts are inconsequential.

Barro (1974, 1978) and James Tobin (1980) discuss a large number of deviations from Ricardian equivalence as various assumptions of the formal theorem are relaxed. Childless couples, alternative models of the bequest motive, corner solutions, imperfect capital markets, and several effects arising from the non-lump sum nature of taxation and from uncertainty receive consideration. Tobin argues that all these effects imply that the replacement of current taxes with a package of debt and concomitant future taxes has a positive effect on aggregate demand. He says nothing, however, about either the relative importance of the various arguments or the quantitative significance of all of them taken together. Barro, on the other hand, while acknowledging deviations from the original hypothesis, concludes that they have indeterminate sign. Hence, he claims, Ricardian equivalence is the appropriate benchmark.

In this paper we examine one particular deviation from Ricardian equivalence (discussed by both Barro and Tobin) and argue that it has both determinate sign and potentially major quantitative significance. Barro writes, "It seems clear that, either in the sense of effects on perceived total wealth, or in the sense of risk composition of household portfolios, the impact of changes in government debt cannot be satisfactorily analyzed without an explicit treatment of the associated tax liabilities" (1974, p. 1115). Taking Barro seriously, we offer such an explicit treatment, noting that taxes are not lump sum, but are positively related to income (indeed, progressively so), and that uncertainty about future income is substantial.

We emphasize the stylized fact (noted by, for example, Robert Lucas, 1977, and documented later in this paper) that variation in individual fortunes is large relative to aggregate uncertainty. A general, though not universal, feature of optimal consumption plans is a precautionary demand for saving (Hayne Leland, 1968). In this case, as long as claims on human capital cannot be traded, a tax cut leads to increased consumption. The reason for this stimulatory effect is that the tax cut provides certain wealth while the future tax increase is contingent upon future income. Taken together, these effects reduce income uncertainty without changing the present value of expected tax payments.[1]

*University of Michigan, Ann Arbor, MI 48109; Harvard University and NBER, Cambridge, MA 02138; The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, respectively. We are grateful to Robert Barro, Olivier Blanchard, Stanley Fischer, Terry Marsh, Torsten Persson, Lawrence Summers, John Taylor, and the referees for helpful comments.

[1]Our examination is a partial equilibrium one, in that we consider only the decision of a consumer in the

The principal result of this paper is that in a stylized but highly suggestive model with plausible estimates of the parameter values, the marginal propensity to consume (*MPC*) out of a tax cut, with associated future income taxes, is likely to be large. Indeed, the *MPC* is in the neighborhood of neo-Keynesian values of the *MPC* that incorporate the life cycle (permanent income) view of consumption, but ignore the future tax liabilities implied by debt finance. Of course, the mechanism we highlight is very different from the usual "bonds are net wealth" channel, since individuals fully perceive all future tax liabilities. In our model, the positive *MPC* is due to the reduction in precautionary saving when the government, by reducing the variance of future income, provides insurance to individuals that is not available in the private market.

Much of this paper is aimed at demonstrating the quantitative importance of the risk-sharing effect on consumption. This effect clearly depends on the nature and amount of individual uncertainty about future labor income. Interpreting the model on Barro's own turf, where operative intergenerational bequests are central, we consider not just uncertainty about one's own income, but uncertainty about the fortunes of future generations as well. Evidence from the available studies of income dynamics suggests that the degree of such uncertainty is likely to be in line with that required for a large marginal propensity to consume.

As is well known, solving for the decision rule of a consumer facing uncertain future income is intractable except in some simple cases. Therefore, to show the potential importance of the risk-sharing effect of a tax cut, we rely on the use of simulations. In particular, we use the technique of stochastic dynamic programming to examine the response of optimal consumption to the income tax cut and future tax increase. Previous authors consider at most the sign of the risk-sharing effect. Through the use of simulation, we are able also to examine its quantitative importance.

Out of necessity, our simulations are highly stylized. The available panel data are not sufficiently detailed to permit estimation and simulation of a complete model of income dynamics with heterogeneous agents. The only tractable strategy is to choose a simple and suggestive specification characterized by a minimal number of parameters, and then to calibrate the model by requiring conformity with the available evidence. The simplicity of our specification allows extensive analysis of the sensitivity of the results to changes in the underlying parameter values.

I. The Model

In this section we show analytically how a tax cut coupled with a future income tax increase can stimulate consumer spending through the precautionary motive for saving. Our development follows that of Louis Kuo Chi Chan (1983), who provides a careful discussion of the importance of missing markets for various deviations from Ricardian equivalence. We examine here just one of these deviations using a two-period model. All individuals in the model are identical *ex ante*. Their labor income in the second period is uncertain and there do not exist markets through which they can insure against this risk.[2] We consider a policy under which the government cuts taxes in the first period, issues bonds to finance the tax cut, and increases income taxes in the second period to repay the debt.

Each individual maximizes expected utility:

$$EU(C_1, C_2), \tag{1}$$

where C_1 = first-period consumption, C_2 = second-period consumption, E = the expectation operator conditional on information

face of a tax cut. Of course, this partial equilibrium effect of a tax cut on consumer spending is a prerequisite for the conventional general equilibrium conclusions.

[2] That is, we exclude markets through which human capital returns can be explicitly traded and also securities with which individual-specific income risk can be hedged.

available in the first period, and $U(\cdot)=$ the von Neumann-Morgenstern utility function.

Before the policy intervention, each individual has first-period labor income μ_1 and second-period labor income $Y_2=\mu_2+\varepsilon$, where ε is a random variable that has zero mean and is uncorrelated across individuals. Although each individual faces uncertainty regarding his future income, there is no aggregate uncertainty.

Each individual can borrow and lend at a risk-free interest rate. Wealth after the first period is

$$W=\mu_1-C_1. \tag{2}$$

Let R be one plus the real interest rate. Second-period consumption is

$$C_2=RW+\mu_2+\varepsilon. \tag{3}$$

In the absence of any government intervention, each individual maximizes expected utility (1) subject to the constraints (2) and (3).

Suppose the government gives each individual a tax cut T in the first period. Since all individuals in the model are identical *ex ante*, the form of the tax cut is irrelevant. Wealth after the first period is

$$W=\mu_1+T-C_1. \tag{2$'$}$$

The government raises taxes to repay the debt in the second period. Suppose it obtains the extra revenue by an increase in a labor income tax.[3] That is, an individual with income Y_2 must pay

$$tY_2 \tag{4}$$

in additional taxes, where $t=$ the tax rate.

[3]Note that capital income is not taxed. If it were, then the policy intervention would lower the after-tax real interest rate, which would also affect consumption. Since our goal is to examine only the risk-sharing effect, we do not include capital taxation. We also do not examine the human capital decision, which in principle is also affected by the policy intervention (Jonathan Eaton and Harvey Rosen, 1980).

The government sets the tax rate t so that the total amount raised equals the debt, which is RT per person in the second period. This government budget constraint requires[4]

$$RT=t(\mu_2), \tag{5}$$

or, equivalently,

$$t=RT/\mu_2. \tag{5$'$}$$

The amount of tax an individual with income Y_2 pays is therefore

$$RT(Y_2/\mu_2). \tag{4$'$}$$

An individual's consumption in the second period is now

$$\begin{aligned} C_2&=RW+\mu_2+\varepsilon-RT(\mu_2+\varepsilon)/\mu_2\\ &=RW+\mu_2-RT+(1-RT/\mu_2)\varepsilon. \end{aligned} \tag{3$'$}$$

Each individual maximizes expected utility (1) subject to the constraints (2′) and (3′). The first-order condition is

$$E[U_1(C_1,C_2)]=RE[U_2(C_1,C_2)]. \tag{6}$$

The three equations (2′), (3′), and (6) jointly determine the three variables C_1, C_2, and W.

We do not solve for the level of consumption C_1, as doing so is intractable except in simple examples. We can solve for the marginal propensity to consume (MPC) out of the tax cut as a function of optimal consumption. By implicitly differentiating the equations (2′), (3′), and (6), we solve for dC_1/dT. We find[5]

$$MPC=\frac{R\,\mathrm{Cov}[(RU_{22}-U_{12}),\varepsilon]}{-\mu_2[EU_{11}-2REU_{12}+R^2EU_{22}]}. \tag{7}$$

[4]More formally, the budget constraint requires that the tax rate times income per capita equals debt per capita. As the size of the population approaches infinity, the tax rate implied by this budget constraint converges in probability to the tax rate implied by (5′).

[5]Equation (7) is parallel to Chan's equation (11).

The *MPC* is not generally zero. A sufficient condition of the *MPC* to be positive is that $RU_{222} - U_{122}$ be uniformly positive.[6] In the additively separable case, the third derivative of the utility function must be positive. In other words, marginal utility must be a convex function of consumption. This condition is even weaker than the condition of nonincreasing absolute risk aversion. Leland and Agnar Sandmo (1970) discuss the more general case and conclude that one should typically expect this condition to hold. Hence, the marginal propensity to consume out of a tax cut is presumptively greater than zero.

A common argument, made by Warren Smith (1969) and Robert Mundell (1971) among others, is that bonds are net wealth because individuals discount the associated future tax liabilities at an interest rate higher than the rate on government bonds. One cannot interpret our analysis in this way. A discount rate for human capital that includes a risk premium and thus exceeds the government bond rate is not sufficient to generate our results. For example, in the case of quadratic utility, optimal consumption decisions display certainty equivalence, despite the risk aversion of the consumer. In this case, the amount the individual would pay today to avoid his tax liabilities is less than their present value computed using the risk-free rate. Nonetheless, the *MPC* out of a tax cut is zero, since the third derivatives of the utility function are zero. The effects of debt and future taxes on the consumption decision cannot be analyzed by reference to any summary wealth statistic.[7]

The positive marginal propensity to consume out of a tax cut relies on the absence of contingent claims markets through which an individual can privately diversify away his individual human capital risk. This assumption appears a reasonable starting point for our analysis, since these contingent claims markets do not in fact appear to exist. Future research could integrate this analysis with an explicit model of missing markets. One way would be to assume that individuals have greater information on the distribution of their future income than is publicly available. It is well-known that the government can provide insurance through mandatory coverage even if adverse selection makes private insurance infeasible.

The model could be made more realistic, as well as greatly more complicated, by including moral hazard.[8] When incentive effects on labor supply are admitted, the increase in insurance achieved through tax cuts may or may not be optimal. Even if government insurance is not optimal, a tax cut that provides insurance will still affect the optimal consumption level of individuals. Following the analysis of Jacques Drèze and Franco Modigliani (1972), one can decompose the risk-sharing effect into an income effect and a substitution effect. We suspect that at the optimal level of government insurance, the marginal deadweight losses exactly balance the income effect, while the substitution effect continues to stimulate current consumption.[9] More generally, we believe that incorporating an explicit model of missing markets will not qualitatively alter the conclusions of this analysis.

The sign of the *MPC* would be affected by altering the risk characteristics of the future tax liabilities. For example, Barro (1974) and Chan point out that if taxes are levied on individuals at random, then a substitution of certain current taxes with random future taxes increases perceived risk and decreases current consumption. We sus-

[6]This result is demonstrated by noting that, for any nondegenerate random variable X and function $F(\cdot)$, if F' is uniformly positive, then $\mathrm{Cov}(X, F(X)) > 0$.

[7]An alternative reason that the future taxes might be discounted at an interest rate higher than that paid on the government debt might be that individuals borrow and lend at different interest rates. Such liquidity constraints are not present in our model.

[8]As the model stands, a 100 percent tax rate with lump sum rebates is optimal. Our purpose, however, is to examine the positive, and not the normative, implications of the risk-sharing effect. Hal Varian (1980) discusses the possible optimality of redistributive taxation as social insurance.

[9]Of course, this will not hold more generally and, in particular, if the marginal tax rate depends on other considerations, such as the need to fund public expenditure.

pect that the risk-creating effect of capricious taxation is empirically less important than the risk-sharing effect of income taxation. Similarly, to the extent that the future tax liabilities fall more heavily on the poor, possibly through reductions in transfer payments, government debt would again be risk-creating. Future research could address these issues more fully.

II. The Extent of Income Uncertainty

The model and the effect we highlight rely on the existence of individual uncertainty regarding future income. Before turning to our simulation results, we examine the evidence on the extent of uncertainty regarding future income. As becomes clear below, this task is not a simple one. In this section, we attempt to use existing analyses of income dynamics to shed some light on the nature of this distribution. The available evidence is consistent with the view that the degree of uncertainty is substantial.

We consider two interpretations of our model. In the first, the uncertainty concerns the income of an individual within his lifetime. In the second interpretation, the uncertainty concerns the performance of future generations of the family. We begin with the former.

A. *Individual Uncertainty*

One interpretation of the model, analogous to many interpretations of overlapping generations models, is that the two periods correspond to the two halves of a single person's life. That is, we can consider each period as corresponding to roughly thirty years. The policy intervention then entails a tax cut during a person's youth coupled with a tax increase during his old age. Under this view, the relevant measure of the uncertainty is that of a young person regarding his income during the second half of his life.

In their analysis of the *Michigan Panel Study of Income Dynamics* (*PSID*), Daniel Hill and Saul Hoffman pose the question, "Does an individual's economic status remain relatively constant over time or is there widespread change in economic standing?" Their conclusion is that "change in status is not only quite common but often quite dramatic as well" (1977, p. 30). In terms of the "income/needs ratio" discussed by Greg Duncan and James Morgan, "less than a quarter of married men were in the same decile position in both 1967 and 1974, about 30 percent changed by one decile, and about 45 percent shifted by two deciles or more" (1977, p. 30).[10]

Another finding from analysis of the *PSID* is that individual incomes are highly vulnerable to disability, which includes medical, psychiatric, and other factors limiting hours of work or precluding work entirely. It is a mistake to conclude that individuals largely insure themselves against income loss from disability. "Even when transfers offset some of the impact, there was a \$3000 to \$5000 a year difference in the family head's income associated with his or her disability" (Morgan, 1980, p. 285).

Robert Hall and Frederic Mishkin (1982), in their study of the sensitivity of consumption to income, provide statistical estimates of the income process that allows us to infer the degree of uncertainty. Using the *PSID* data, they first use regression to correct family income for life cycle and other demographic effects. They then divide the residual into a lifetime component, which follows a random walk, and a transitory component, which follows a second-order moving-average process. Over a forecast horizon of thirty years, the variance of the lifetime component far exceeds the variance of the transitory component. Hall and Mishkin report that the annual innovation to the lifetime component has a standard deviation of about \$1200. The standard error of a forecast over a thirty-year horizon is thus \$6600. Since the median family income during their time period (1972) was roughly \$12,000, the implied coefficient of variation is 0.55.[11]

[10] Hill and Hoffman (p. 33) also report that the largest share of variation in the income/needs ratio comes from income rather than needs.

[11] Alternatively, one might look at the uncertainty concerning average income in the second half (30 years) of life. If the individual is half way through the first

The uncertainty in our model is individual rather than aggregate. This assumption is important, since the government cannot provide insurance against aggregate shocks to income. It is, however, also empirically valid. Hall and Mishkin report that the "overwhelming bulk of movements in income that give rise to our inference from the data are unrelated to the behavior of the national economy; most are probably highly personal" (p. 480). Thus, the observed degree of uncertainty is correctly interpreted as a measure of individual rather than aggregate risk.

While the results from these studies are consistent with the view that there is substantial income uncertainty, this interpretation is certainly not free of problems. Not all of the measured variation reflects true uncertainty about lifetime earnings. Measurement error is one problem that arises when using panel data.[12] In the Hall and Mishkin study, however, the measurement error is likely to be included in the transitory component of income and thus should not affect the estimated conditional variance of the lifetime component.

A second problem of these studies is that the individual agents may have greater information than the econometrican. That is, what is "news" to the econometrican may not be news to the individual. Benjamin Eden and Ariel Pakes (1981) develop a methodology that can deal with this problem, as well as the measurement error problem, using the restrictions imposed by the permanent income hypothesis (Hall, 1978). In particular, only genuine news about permanent income should affect consumption. Unfortunately, the data Eden and Pakes use are not of high quality, resulting in a large confidence interval for the variance of the innovation in permanent income. Nevertheless, this methodology may provide an avenue through which future research can resolve this difficulty.

B. *Intergenerational Uncertainty*

A second interpretation of the model is that the two periods represent two generations. The relevant measure of uncertainty is that of a person forecasting the income of his child. Perhaps surprisingly, it is easier to glean evidence on the conditional distributions of sons' and grandsons' incomes than on the conditional distribution of own income. The distribution of a descendant's income presumably depends on a small number of identifiable characteristics.

A classic reference for the distribution of earnings conditional on family background, educational attainment, and occupational status is Christopher Jencks (1972). Among his striking findings are:

1) Upper-middle-class parents are unable to ensure that their children will maintain their privileged position. Among men born into the most affluent fifth of the population, only 40 percent will be in this top quintile as adults (p. 215).

2) Correlation between parents' and son's permanent incomes is only about 0.3 (p. 236).

3) Family background explains about 15 percent of the variation in earnings. The earnings of brothers raised in the same home would vary radically. "In 1968, for example, if we had compared random pairs of individuals, we would have found that their earnings differed by an average of about $6,200. If we had had data on brothers, our best guess is that they would have differed by at least $5,600." If the earnings of the general population exhibited only the degree of inequality characteristic of brothers, the best-paid fifth of all male workers would still earn six times the pay of the lowest quintile (pp. 219–20).

4) "Neither family background, cognitive skill, educational attainment, nor occupational status explains much of the variation in men's incomes. Indeed, when we

(30-year) period and assuming income follows a random walk (i.e., ignoring the transitory component), the coefficient of variation of the forecast of second period average income is 0.50, close to the 0.55 of the text. Thomas MaCurdy (1982) estimates an earnings equation with a different stochastic specification using the same data. Calculations based on his estimates imply a similar coefficient of variation.

[12] See, for example, Joseph Altonji and Aloysius Siow (1984).

compare men who are identical in all these respects, we find only 12 to 15 percent less inequality than among random individuals" (p. 226).

The following table compares several parameters of the conditional distribution of earnings given father's education and occupational status with the corresponding parameters of the unconditional distribution. The underlying data are earnings of full-time, year-round, male workers in 1968 (p. 236).

	Unconditional Distribution	Conditional Distribution Given Father's Education and Occupational Status
Standard Deviation	\$5,508	\$5,232
Ratio of Mean of Top 5th to Mean of Bottom 5th	7.7	6.5

Jencks interprets these numbers as evidence indicating a large random component in the determination of lifetime earnings. In summary, "luck has far more influence on income than successful people admit" (p. 227).

Some studies, such as that of John Brittain (1977), criticize Jencks on a variety of grounds: for not using actual data on brothers, for underestimating the correlation of income within families, and for jumping to excessively strong conclusions given his evidence. But, as the sophisticated studies in Paul Taubman's (1977) volume indicate, repeating Jencks's exercise with actual data on brothers and with more advanced statistical techniques leads to almost identical conclusions. For instance, Michael Olnick writes, "The average difference between brothers on earnings is 87 percent as large as the difference between random individuals" (1977, p. 137). Thus no parent can feel assured of even roughly predicting his children's future earnings.

III. Two-Period Simulations

The theory shows that, under plausible conditions, the marginal propensity to consume out of a tax cut is positive because of the risk-sharing effect. Examination of the degree of income uncertainty suggests that human capital returns are indeed risky and undiversifiable through contingent claims markets. We now turn to the question of whether the risk-sharing effect is quantitatively large. We answer this question by simulating the consumer's optimization problem for reasonable parameter values. The simulation method is explained in the Appendix.

A major difficulty in attempting to quantify the risk-sharing effect arises when deciding on an appropriate way to use the limited, though suggestive, evidence on income uncertainty reviewed in Section II. In principle, one approach would be to construct a multipoint distribution that would mimic the uncertainty regarding lifetime income faced by a typical family. We do not adopt this approach because the available evidence is much too scanty to allow us to pin down such a distribution. Instead, we choose a simple, symmetric three-point distribution, which is characterized by two parameters. We then calibrate these parameters by requiring that the implied coefficient of variation (standard deviation divided by mean) be consistent with the results of Hall and Mishkin and others.

We assume throughout that the utility function is time-separable. We begin with two-period simulations. As discussed above, one can interpret the simulations in two ways. The first interpretation is that each period represents one-half of a single life.

During the first half of the individual's life, he earns \$100. During the second half, he also expects to earn \$100. This latter income, however, is uncertain. We assume that second-period income follows the distribution:

$$
\begin{aligned}
Y_2 = (1-x)100 &\quad \text{with probability } p, \\
100 &\quad \text{with probability } 1-2p, \\
(1+x)100 &\quad \text{with probability } p.
\end{aligned}
$$

With some probability p, his income falls below its mean value of 100. One can view this unlucky event as a variety of possible

outcomes. As discussed above, the degree of income uncertainty is great for the typical individual. The individual could become disabled, losing much of his earning power. The individual might lose his job in a high-paying industry because of technological innovation or foreign competition. Or he simply could turn out less successful in his chosen occupation than he anticipated. The first outcome in our three-point distribution represents the "bad" event which, although possibly unlikely, may be sufficiently worrisome to generate a precautionary demand for saving.

The distribution of the individual's future income is symmetric, so that there is also a probability p of an extraordinarily good event. Individuals find themselves more successful in their careers than they expected. This sort of event is represented in the third outcome in the above distribution.

The second interpretation of the model is that the first period represents an individual's life, while the second represents the life of his child. Under this view, the individual is relatively certain of his own lifetime income, but his child's lifetime income is unknown. (Indeed, his child may not even be born yet.) For concreteness, we discuss the simulation as if it were two periods of a single life.

We consider a tax cut that gives the individual T in the first period along with a contingent tax liability in the second period.[13] In the bad state, the individual pays no tax. In the two other states, he pays a tax proportional to his income in excess of the floor income $(1-x)100$. In expectation, the present value of his tax liability equals his tax cut.[14]

The policy intervention we consider is a *marginal* tax change for an economy in which taxes and transfers already exist. Therefore, Y_2 is income net of these existing taxes and transfers. The income floor of $(1-x)100$ is possibly due to existing government programs. We assume that this income floor is not affected by the policy intervention.[15]

Our three-point distribution exhibits a coefficient of variation (σ) equal to $x(2p)^{1/2}$. As a conservative summary of the estimated coefficient of variation from the empirical literature, we take σ to be at least 1/3 but no larger than 1/2. Of course, many combinations of p and x are consistent with a given value of σ. Moreover, the marginal propensity to consume out of a tax cut depends on the entire distribution of income, not just its second moment. It is therefore important to examine the various combinations of the two parameters to ensure that any conclusion is robust.

Table 1 presents the results of the simulations for two scenarios. The real interest rate is zero ($R=1$) and the utility function of the consumer is additively separable through time with no time preference. The single-period utility function exhibits constant relative risk aversion.[16] For the results in panel A, the coefficient of relative risk aversion is one, while for the results in panel B, the coefficient of relative risk aversion is three.[17] The region for which $1/3 \leq \sigma \leq 1/2$ is marked in each panel.

Implicit in much neo-Keynesian analysis of tax cuts, such as that of Alan Blinder (1981), are two assumptions. First, consumers set their consumption in proportion to the present value of expected income. In other words, their behavior is based on the life cycle (permanent income) theory of con-

[13] The *MPC*s reported for these two-period examples are for an infinitesimal T; these are very close to the *MPC*s calculated for a T of 5 percent of first-period income.

[14] That is, $RT=E[t(Y_2-(1-x)100)]$.

[15] Alternatively, one could assume that the tax increase is strictly proportional, rather than progressive. In this case, the *MPC* is exactly the product of x and the *MPC* as we compute it.

[16] One property of this iso-elastic utility function is that a proportionate change in income, wealth, and the distribution of future income does not affect the marginal propensity to consume. Heterogeneity regarding current (certain) wealth relative to future (uncertain) income is ignored here but is discussed in Zeldes (1986). Zeldes's simulations suggest that the *MPC* is decreasing in the level of certain wealth relative to uncertain income.

[17] Recent studies that estimate the coefficient of relative risk aversion find values in this range. See, for example, Lars Hansen and Kenneth Singleton (1983) or Mankiw (1985).

TABLE 1—THE MARGINAL PROPENSITY TO CONSUME

	$x=1/4$	$x=1/2$	$x=3/4$	$x=1$
A. Logarithmic Utility[a]				
$p=1/128$	0.50	0.50	0.52	0.76
	[0.00]	[0.01]	[0.06]	[0.56]
$p=1/32$	0.50	0.51	0.55	0.73
	[0.02]	[0.05]	[0.16]	[0.56]
$p=1/8$	0.51	0.53	0.58	0.68
	[0.07]	[0.17]	[0.34]	[0.57]
$p=1/4$	0.51	0.54	0.59	0.65
	[0.13]	[0.28]	[0.44]	[0.58]
$p=1/2$	0.52	0.55	0.58	0.61
	[0.24]	[0.41]	[0.52]	[0.58]
B. Relative Risk Aversion of Three[b]				
$p=1/128$	0.50	0.51	0.61	0.78
	[0.01]	[0.06]	[0.34]	[0.73]
$p=1/32$	0.50	0.53	0.63	0.73
	[0.04]	[0.17]	[0.48]	[0.69]
$p=1/8$	0.51	0.55	0.61	0.66
	[0.14]	[0.36]	[0.55]	[0.64]
$p=1/4$	0.52	0.56	0.59	0.62
	[0.25]	[0.45]	[0.56]	[0.61]
$p=1/2$	0.52	0.55	0.56	0.57
	[0.39]	[0.52]	[0.55]	[0.57]

Notes: The top number is the *MPC* out of tax cut alone; the number below in brackets is the *MPC* out of a tax cut coupled with a future income tax increase.

[a]Assumptions: $U(C_1, C_2) = \log(C_1) + \log(C_2)$.

[b]Assumptions: $U(C_1, C_2) = (C_1^{1-A})/(1-A) + (C_2^{1-A})/(1-A)$; $A = 3$.

$$R = 1.0;\; Y_1 = 100;\; Y_2 = \begin{cases} (1-x)100 & \text{with prob. } p \\ 100 & \text{with prob. } 1-2p \\ (1+x)100 & \text{with prob. } p \end{cases}$$

sumption and on certainty equivalence. Second, the future tax liabilities implied by debt finance are ignored, presumably because they fall on some future generation. Under these two assumptions, the *MPC* out of a tax cut in a two-period model with no discounting is 0.5. Thus, we take 0.5 to be the benchmark "Keynesian" estimate.

A. *Excess Sensitivity*

The first important observation is that consumption exhibits "excess sensitivity" to current income. Much work on consumption, not only that of Blinder on tax cuts, but also that of Marjorie Flavin (1981), Hall and Mishkin, and Ben Bernanke (1985), rests on the assumption that optimal consumption exhibits certainty equivalence. In this case, one need look only at the first moment of income to determine the optimal level of consumption. As pointed out above, certainty equivalence implies an *MPC* out of wealth of 0.5 in our two-period example.

As Zeldes shows, utility functions with positive third derivatives can exhibit "excess sensitivity," even though consumption is set optimally and there are no borrowing constraints.[18] The top numbers in Table 1 show the *MPC* out of a tax cut with no associated future tax increase for various degrees of uncertainty. These *MPC*s are greater than 0.5, the value one would obtain assuming certainty equivalence.

[18]Another and very different explanation of excess sensitivity is suggested by Ron Michener (1984): in general equilibrium, rates of return may vary to make consumption more closely track income.

B. *A Bird in the Hand*

The numbers in brackets in Table 1 are the *MPC*s out of a tax cut coupled with a future income tax increase. The tax change has no effect on the individual's permanent income as defined by, for example, Flavin. Yet the tax change can often have very large effects on consumption.

If we assume $1/2 \leq \sigma \leq 1/3$, then the *MPC* is never smaller than 0.28 with logarithmic utility and 0.45 with relative risk aversion of three. The individual's marginal propensity to consume out of a one dollar tax cut is high, even though he will, on average, have to repay the dollar to the government in the second period.[19] The consumer is Ricardian in taking into account the future tax liabilities implied by debt finance, and is Keynesian in increasing his spending in response to the tax cut.

A comparison of the top and bottom numbers demonstrates the importance of the future tax increase as a factor mitigating the stimulative effect of the tax cut. For distributions with little uncertainty (small x and p), the tax increase almost fully eliminates the effect of the tax cut on spending. For distributions with $1/3 \leq \sigma \leq 1/2$, which appear to fit the stylized facts we discuss above, the future tax increase provides only a small mitigating effect. Because the individual pays no taxes if "times are bad" for him, the future liability has little effect on current consumption. The tax cut, like a bird in the hand, stimulates spending, despite the contingent tax increase. Indeed, a naive observer might wonder if the consumer simply ignores his future tax liability altogether.

C. *Unlikely and Unlucky Events*

It is particularly interesting to compare the two *MPC*s for the $x=1$ column. With these distributions, there is a small but nonzero probability of zero income in the second period. In this unlucky event, the individual consumes only what he saved from the first period.

The *MPC* out of a tax cut, along with the future income tax increase, is very large for all these distributions. Even if the unlucky event is unlikely ($p=1/128$), the uncertainty is sufficient to generate a large *MPC*: 0.56 in panel A and 0.73 in panel B. Remember that if p were equal to zero, the *MPC* would also be zero. It appears that consumption and saving behavior can be greatly affected by small probability events.

One might argue that a second-period income of zero is unrealistic, since various institutions in society provide a floor on income. Although the existence of such a floor is undeniable, it is also true that there is some consumption level below which survival is impossible. Suppose that society provides a floor on income at the survival level, C_s, and that utility is defined in excess of this survival level as

$$U(C) = (C - C_s)^{1-A}/(1-A).$$

In this case, the results in the $x=1$ column continue to apply, regardless of the level of the income floor.

D. *The Rates of Interest and Time Preference*

In the above simulations, we assume that the real interest rate between the two periods is zero and that individuals do not discount future relative to present utility. Table 2 presents results that relax these assumptions. Since the two periods represent two halves of a single life, we use a real interest rate of 50 percent and a comparable discount rate. We see that a higher real interest rate lowers the *MPC*s, while a higher rate of time preference raises the *MPC*s. Our primary conclusion—that a tax cut can have a large impact on consumer spending despite the future tax liabilities—is not affected by alternative rates of interest and time preference.

E. *Growth in Income*

The simulation results above assume no growth in expected income. In reality, the growth of per capita real income is not zero

[19] The optimal level of saving in this example is 7.5 percent of first-period income.

TABLE 2—THE MARGINAL PROPENSITY TO CONSUME: ALTERNATIVE RATES OF INTEREST AND TIME PREFERENCE[a]

	$x=1/4$	$x=1/2$	$x=3/4$	$x=1$
$R=\beta=1.0$	0.51	0.54	0.59	0.65
	[0.13]	[0.28]	[0.44]	[0.58]
$R=\beta^{-1}=1.5$	0.61	0.63	0.68	0.73
	[0.15]	[0.32]	[0.49]	[0.63]
$R=1.0$	0.61	0.66	0.72	0.77
$\beta^{-1}=1.5$	[0.19]	[0.41]	[0.60]	[0.73]
$R=1.5$	0.51	0.52	0.56	0.61
$\beta^{-1}=1.0$	[0.10]	[0.21]	[0.34]	[0.47]

Notes: See Table 1.
[a]Assumptions: $U(C_1,C_2)=\log(C_1)+\beta\log(C_2)$.

$$Y_1=100;\ Y_2=\begin{cases}(1-x)100 & \text{with prob. } 1/4\\ 100 & \text{with prob. } 1/2\\ (1+x)100 & \text{with prob. } 1/4\end{cases}$$

but has averaged about 2.2 percent annually since 1960. Taking into account such growth strengthens our conclusions by making a higher fraction of lifetime income uncertain. In particular, if we scale up each of the possible outcomes by a constant growth factor $(1+g)$ and assume the same tax structure (proportional to income in excess of the floor income), the new *MPC* is the same as if we had increased the parameter x by a factor of $(2+2g)/(2+g)$. For a value of $1+g$ of $1.92=(1.022)^{30}$, the *MPC* for any x_0 is the same as the *MPC* without growth for $x=1.31x_0$. If $p=1/4, x=1/2, g=0.92$, the *MPC* out of a tax cut with logarithmic utility would be 0.38, as opposed to a *MPC* of 0.28 in an economy without growth.

F. *A Multipoint Income Distribution*

As a final two-period simulation, we try a multipoint income distribution. Again, there is no discounting of any sort. We consider the two periods as two generations. The father earns \$100 with certainty in the first period. The son also expects to earn \$100. We base the distribution for the son on the distribution of the earnings of full-time, year-round male workers in 1970, as reported by Jencks (p. 213). In particular, the son's income distribution is[20]

\$12	with prob. 0.047	\$105	with prob. 0.179
35	0.082	146	0.195
58	0.171	204	0.063
82	0.244	350	0.019

We compute the *MPC* for the utility function exhibiting constant relative risk aversion of three. The *MPC* out of a tax cut with no future tax liability is 0.60, while the *MPC* out of a tax cut with a future proportional income tax increase is 0.41.[21] This latter value of the *MPC* out of a tax cut is closer to the Keynesian benchmark of 0.5, than to the Ricardian benchmark of zero.

To test the robustness of our result to alternative forms of the utility function, we also compute the *MPC* for this multipoint distribution using a constant *absolute* risk-

[20] This distribution, which has a coefficient of variation of about 0.6, overestimates the uncertainty by including some transitory and life cycle variation in income, but underestimates the uncertainty by excluding all disability and chronic unemployment.

[21] The level of saving in this example is 23 percent of first-period income. This finding suggests that the precautionary motive for saving may be an important explanation for the high level of bequests reported by Laurence Kotlikoff and Lawrence Summers (1981). Interestingly, the family in this example would pay 36 percent of its first-period income to eliminate second-period income uncertainty entirely (keeping the mean constant).

aversion utility function. We choose the coefficient of absolute risk aversion so that the coefficient of relative risk aversion at the mean of second-period income is equal to three (the value we use above).[22] In this case, the MPC out of a tax cut alone is 0.50, while the MPC out of a tax cut with the future tax increase is 0.24. Thus, the risk-sharing effect continues to be important with this alternative specification of preferences.

IV. Multiperiod Simulations

In this section, we investigate how our results are affected by extending the number of periods in the model.[23] In particular, we explore how the MPC out of a tax cut is affected by the horizon over which the debt is to be repaid. The model includes five periods and there is no discounting of any sort. Each period here represents a generation. Income is independently and identically distributed in each generation. Because family characteristics have little value in predicting earnings, it seems a reasonable approximation to assume that the uncertainty about the fate of one's grandchildren is not greater than the uncertainty about one's children.

In a world of the type Barro describes, the MPC out of tax cut equals zero regardless of the timing of the corresponding tax increase. In a certainty or certainty equivalent model with no future taxes, the MPC equals 0.2. Thus, 0.2 is the benchmark "Keynesian" estimate.[24] Table 3 presents the MPCs implied by a utility function with constant relative risk aversion of three and no discounting of any sort. The MPC for the case in which there is no future tax increase can exceed 0.2 by large amounts. Again, this effect is the "excess sensitivity" of consumption to current income.

TABLE 3—THE MARGINAL PROPENSITY TO CONSUME: ALTERNATIVE DEBT REPAYMENT HORIZONS[a]

Taxes Repaid In Period:	$x=1/2$	$x=3/4$	$x=1$
2	0.03	0.10	0.35
3	0.04	0.15	0.39
4	0.07	0.20	0.41
5	0.14	0.25	0.42
Never	0.22	0.27	0.42

Note: This table shows the MPC out of a first-period tax cut, varying the period during which the future tax increase occurs.

[a]Assumptions: $U(C_1,C_2,C_3,C_4,C_5)=\sum_1^5 \frac{C_i^{1-A}}{1-A}$; $A=3$.

$R=1.0$; $Y_1=100$;
$Y_i=(1-x)100$ with prob. $1/8$ $i=2,3,4,5$
100 with prob. $3/4$
$(1+x)100$ with prob. $1/8$

The results that include the future tax liability are dramatic. We find that the repayment horizon is critical to the effect of the tax cut on consumption. The farther in the future is the tax increase, the higher is the MPC out of the current tax cut. Risk sharing in a later period has greater effect on consumption than risk sharing in an early period. This result is due to the fact that a tax increase in a later period implies an earlier resolution of uncertainty. Indeed, if the taxes are not raised until period 5, the MPCs are almost as large as if the taxes are not raised at all. Consumers have MPCs that are very close to being Keynesian, even though they fully incorporate all future tax liabilities in their plans. Indeed, the MPCs we find sometimes exceed the Keynesian benchmark of 0.2.[25]

[22]Thus, the utility function is $-\exp(-aC)$, and a is 3/100.

[23]It is not the case that increasing the number of time periods diversifies away identical and independently distributed income. Numerical examples in Zeldes demonstrate that, for a given income process and initial wealth, precautionary saving does not decrease when the number of periods increases. This result is closely related to Paul Samuelson's (1963) discussion of repeated gambles.

[24]The low value of the Keynesian benchmark is in part due to the absence of any discounting in our example.

[25]We also tried an intervention in which the government announces a tax cut in period 1 to go into effect in period 2, coupled with a tax increase in period 5. The MPCs were 0.03 for $x=1$, 0.13 for $x=3/4$, and 0.10

The results in Table 3 assume that income is independently distributed in each period. More realistically, income might be modeled as containing both permanent and transitory components. In this case, the uncertainty regarding income in later periods is greater than the uncertainty regarding income in earlier periods. The length of the repayment horizon would be even more important in this case. The results in Table 3 might thus understate the importance of the repayment horizon.

V. Conclusion

In this paper we examine the interaction between individual income uncertainty and income taxation in the face of a debt-financed tax cut. Under plausible assumptions regarding preferences toward risk, the marginal propensity to consume out of a tax cut, coupled with a future income tax increase, is positive because of an increase in risk sharing. An examination of the degree of income uncertainty suggests that this uncertainty is substantial, indicating that the risk-sharing effect may be important. Numerical simulations show that this effect is potentially large. Indeed, the *MPC* out of a tax cut, coupled with a future income tax increase, appears closer to the Keynesian value that ignores the future taxes than to the Ricardian value that treats the future taxes as if they were lump sum.

The final question to address is whether the risk-sharing effect provides an intuitively appealing reason to believe that tax cuts stimulate consumer spending. To hone one's intuition, it is useful to envision two policy interventions. In both, the government gives a tax cut of $2000 now to every living individual and will raise the $2000 and accumulated interest by a tax on the next generation in thirty years.

In the first intervention, which probably corresponds best to what we experience, the future tax increase takes the form of a marginal change in income tax rates. This change will raise the necessary revenue in total, but not necessarily the same from each individual's child.

In the second intervention, the government levies a lump sum tax. That is, from each child, the government will collect the $2000 and accumulated interest, regardless of how impoverished the child happens to be and regardless of how dire the consequences. While it is difficult to envision such a tax increase, it seems clear that this prospect, if credible, would cause more concern among parents than a mere change in income tax rates. In particular, a prospective lump sum tax would plausibly seem to call forth greater saving to meet the future liability. While Ricardian equivalence may be the appropriate benchmark for a world in which taxes are lump sum, it is probably not the appropriate benchmark for the world in which we live.

Appendix: Simulation Method

For the two-period examples, equation (7) gives the analytical expression for the *MPC* out of a tax cut in period one.[26] The right-hand side of equation (7), however, must be evaluated at the optimal choice of consumption, which in general cannot be calculated analytically. We therefore use numerical methods to calculate the optimal level of consumption and then use this value in equation (7) to arrive at the *MPC*s.

In the multiperiod examples, we do not use an analytic expression to compute the *MPC*s. We use numerical methods to calculate the optimal level of consumption in each example both before and after the tax cut. The *MPC* out of the tax cut is the difference in consumption divided by the size of the tax cut.

The technique used to calculate the optimal consumption levels is stochastic dynam-

for $x = 1/2$. For a tax cut effective in periods 1 and 2, coupled with a tax increase in periods 4 and 5, the *MPC*s are 0.46 for $x = 1$, 0.32 for $x = 3/4$, and 0.16 for $x = 1/2$.

[26] For most of the simulations, the tax increase is not proportional to second-period income. In these cases, an expression analogous to (7) is derived.

ic programming.[27] First, the problem is formulated as a stochastic control problem with one state variable (current wealth), one control variable (consumption), and one disturbance (income).

The state space is made discrete using a technique suggested by Dimitri Bertsekas (1976). The first step is to determine the upper and lower bounds for wealth (W). If there is a positive floor on income in future periods, we can, without loss of generality, redefine income as the amount in excess of the floor and redefine wealth to include the present value of all future income floors. In this new problem, there is positive probability of receiving an income of zero in each period. Since we use utility functions for which $U'(0) =$ infinity, we know that individuals will always carry positive "wealth," that is, they will never borrow against risky labor income. For this new problem, then, W must always be greater than zero, so we choose $W_{\min} = 0$.

We next need to determine an upper bound for wealth, that is, a $W_{\max}$ that wealth never exceeds at any point during life. One choice would be the wealth that would be accumulated if income turned out to be its maximum in each period and consumption equalled zero in each period.

Next, we divide the range $(W_{\min}, W_{\max})$ into G equal intervals. For our two-period examples, we use $G = 1000$. In other words, the state variable W can take on any of the 1000 discrete values between $W_{\min}$ and $W_{\max}$. For the last period of life, optimal consumption is equal to wealth, and the value function is equal to the utility function. In all prior periods, the computer searches, for each level of the state variable, for the choice of consumption that maximizes the sum of current utility and the discounted expected value of next period's value function.[28]

[27] Zeldes describes this technique in more detail, and uses the technique to investigate some of the properties of optimal consumption in the presence of nontraded labor income.

[28] While there are simpler methods for calculating the two-period results (such as numerically approximating the solution to the single Euler equation), the advantage of this method is that the same technique can be used regardless of the number of the periods in the model.

While the numbers are an approximation to the actual solution, we can make the approximation errors arbitrarily small by narrowing the width of the grid used for the discretization (see Bertsekas). We tested our grid against some simple examples that can be solved analytically. The results were very close. We believe that our calculated *MPC*s are accurate to ± 0.03.

The examples used here took little computer time to run. Yet as the number of time periods increases, $W_{\max}$ must also increase, implying that the required memory and CPU time increases dramatically.

REFERENCES

Altonji, Joseph G. and Siow, Aloysius, "Testing the Response of Consumption to Income Changes with (Noisy) Panel Data," Columbia University, 1984.

Barro, Robert J., "Are Government Bonds Net Wealth?," *Journal of Political Economy*, November/December 1974, *82*, 1095–117.

______, "Public Debt and Taxes," in M. J. Boskin, ed., *Federal Tax Reform*, San Francisco: Institute for Contemporary Studies, 1978.

Bernanke, Ben, "Adjustment Costs, Durables and Aggregate Consumption," *Journal of Monetary Economics*, January 1985, *15*, 41–68.

Bertsekas, Dimitri, *Dynamic Programming and Stochastic Control*, New York: Academic Press, 1976.

Blinder, Alan S., "Temporary Income Taxes and Consumer Spending," *Journal of Political Economy*, February 1981, *89*, 26–53.

Brittain, John A., *The Inheritance of Economic Status*, Washington: The Brookings Institution, 1977.

Chan, Louis Kuo Chi, "Uncertainty and the Neutrality of Government Financing Policy," *Journal of Monetary Economics*, May 1983, *11*, 351–72.

Drèze, Jacques H. and Modigliani, Franco,

"Consumption Decisions under Uncertainty," *Journal of Economic Theory*, December 1972, *5*, 308–35.

Duncan, Greg J. and Morgan, James N., "An Overview of Part I Findings," in their *Five Thousand American Families*, Vol. V, Ann Arbor: University of Michigan, ISR, 1977.

Eaton, Jonathan and Rosen, Harvey S., "Taxation, Human Capital, and Uncertainty," *American Economic Review*, December 1980, *70*, 705–15.

Eden, Benjamin and Pakes, Ariel, "On Measuring the Variance-Age Profile of Lifetime Earnings," *Review of Economic Studies*, July 1981, *48*, 385–94.

Flavin, Marjorie A., "The Adjustment of Consumption to Changing Expectations of Future Income," *Journal of Political Economy*, October 1981, *89*, 974–1009.

Hall, Robert E., "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence," *Journal of Political Economy*, December 1978, *86*, 971–87.

______ **and Mishkin, Frederic S.**, "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households," *Econometrica*, March 1982, *50*, 461–81.

Hansen, Lars Peter and Singleton, Kenneth, "Stochastic Consumption, Risk Aversion and the Temporal Behavior of Asset Returns," *Journal of Political Economy*, April 1983, *91*, 249–65.

Hill, Daniel and Hoffman, Saul, "Husbands and Wives," in G. J. Duncan and J. N. Morgan, eds., *Five Thousand American Families*, Vol. V, Ann Arbor: University of Michigan, ISR, 1977.

Jencks, Christopher, *Inequality*, New York: Basic Books, 1972.

Kotlikoff, Laurence J. and Summers, Lawrence H., "The Importance of Intergenerational Transfers in Aggregate Capital Accumulation," *Journal of Political Economy*, August 1981, *89*, 706–32.

Leland, Hayne E., "Saving and Uncertainty: The Precautionary Demand for Saving," *Quarterly Journal of Economics*, August 1968, *82*, 465–73.

Lucas, Robert E., Jr., "Understanding Business Cycles," in Karl Brunner and Alan Meltzer, eds., *Stabilization of the Domestic and International Economy*, Vol. V, Carnegie-Rochester Series on Public Policy, *Journal of Monetary Economics*, Suppl. 1977, 7–29.

MaCurdy, Thomas E., "The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis," *Journal of Econometrics*, January 1982, *18*, 83–114.

Mankiw, N. Gregory, "Consumer Durables and the Real Interest Rate," *Review of Economics and Statistics*, August 1985, *67*, 353–62.

Michener, Ron, "Permanent Income in General Equilibrium," *Journal of Monetary Economics*, May 1984, *13*, 297–306.

Morgan, James N., "Occupational Disability and Its Economic Correlates," in G. J. Duncan and his *Five Thousand American Families*, Vol. VII, Ann Arbor: University of Michigan, ISR, 1980.

Mundell, Robert, *Monetary Theory*, Pacific Palisades: Goodyear, 1971.

Olnick, Michael R., "On the Use of Sibling Data to Estimate the Effects of Family Background, Cognitive Skills, and Schooling: Results from the Kalamazoo Brothers Study," in P. Taubman, ed., *Kinometrics: Determinants of Socioeconomic Success within and between Families*, Amsterdam: North-Holland, 1977.

Samuelson, Paul, "Risk and Uncertainty: A Fallacy of Large Numbers," *Scientia*, April-May 1963; reprinted in J. E. Stiglitz, ed., *The Collected Scientific Papers of Paul A. Samuelson*, Vol. I, Cambridge: MIT Press, 1966.

Sandmo, Agnar, "The Effect of Uncertainty on Saving Decisions," *Review of Economic Studies*, July 1970, *37*, 353–60.

Smith, Warren, "A Neo-Keynesian View of Monetary Policy," in *Controlling Monetary Aggregates*, Boston: Federal Reserve Bank, 1969.

Taubman, Paul, *Kinometrics: Determinants of Socioeconomic Success within and between Families*, Amsterdam: North-Holland, 1977.

Tobin, James, *Asset Accumulation and Economic Activity*, Chicago: University of Chicago Press, 1980.

Varian, Hal R., "Redistributive Taxation as Social Insurance," *Journal of Public Economics*, August 1980, *14*, 49–68.

Zeldes, Stephen P., "Optimal Consumption with Stochastic Income: Deviations from Certainty Equivalence," University of Pennsylvania, 1986.

In Defense of Base Drift

By CARL E. WALSH*

By using the actual money supply as the base for its target ranges, the Federal Reserve impounds past target misses into the level of its new target ranges. This practice of allowing "base drift" has often been criticized. A simple aggregate model is used to derive the optimal degree of base drift. It is shown that some drift will be optimal if income and/or velocity are nonstationary.

Each February, target growth rates for various monetary aggregates are announced by the Federal Reserve's Open Market Committee (FOMC). These growth rates, which apply from the fourth quarter of the previous year to the fourth quarter of the current year, have been publicly announced since the passage of House Concurrent Resolution 133 in 1975.[1] It has been the practice of the FOMC to use the fourth-quarter values of each aggregate as the base from which the following year's growth path is calculated. This procedure results in what is known as base drift: if the actual value of an aggregate in the fourth quarter differs from that year's target, this deviation is made permanent by being impounded in the level of the aggregate along the following year's growth path. The FOMC could, of course, compensate for any such base drift by adjusting the target growth rate downward if drift had been positive and upward if it had been negative. The evidence, however, suggests that this has not been done.[2]

Because the FOMC establishes target ranges rather than a single-target growth rate, one estimate of the amount of base drift allowed by the FOMC would be the difference between the actual fourth-quarter value of an aggregate such as $M1$ and the fourth-quarter value implied by the midpoint of that year's target growth range for $M1$. This measure of drift is given in Table 1.[3] As the table shows, the actual fourth-quarter level was above the midpoint of the target range for seven of the ten years from 1976 to 1985, and for every year during the period 1976–83 except 1981. These above-target levels then formed the bases from which the subsequent target ranges were calculated, leading to a permanently higher level of $M1$ and, over the period as a whole, a higher growth rate as well. The opposite occurred in 1984 with $M1$ below the midpoint of the target range in the fourth quarter of that year. This low $M1$ was then used as a base for the 1985 growth path. In July of 1985, with $M1$ running well above the target ranges established in February, the FOMC

*Federal Reserve Bank of San Francisco, Economic Research, 101 Market St., San Francisco, CA 94105, and National Bureau of Economic Research. This paper was written while I was at Princeton University. I would like to thank, without implicating, Alan Blinder, Peter Hartley, Mike Hutchison, Jan Loeys, and Jim Wilcox for helpful comments. This paper has benefited greatly from the comments and suggestions of an anonymous referee and John Taylor. Any opinions expressed are my own and do not necessarily reflect the views of the Federal Reserve Bank of San Francisco or the Federal Reserve System.

[1] Preliminary target ranges for the following year are announced in July. In 1979, 1983, and 1985, the ranges announced in February were modified in July. For an interesting discussion of the setting of the target ranges in 1983 and 1984, see R. W. Hafer (1985).

[2] For example, the correlation between base drift in year t (see Table 1 below) and the midpoint of the growth rate range established for year $t+1$ is positive. Thus, above-average base drift has tended to be followed not by lower target ranges, but by above-average target growth ranges.

[3] The numbers in Table 1 differ from those in chart 1–4 in the *Economic Report of the President*, (1985, p. 53) because the Council's chart reflects estimated adjustments to $M1$ due to financial deregulation. These shift adjustments are discussed in Section II.

TABLE 1—ESTIMATED BASE DRIFT OF $M1$[a]
(Billions of Dollars)

	Midpoint of Target Range for Fourth Quarter (1)	Actual Fourth-Quarter Value (2)	Base Drift (2)–(1)
1976	306.8	305.6	−1.2
1977	322.4	329.5	7.1
1978	346.8	353.9	7.1
1979	367.0[b]	371.2	4.2
1980	408.2	416.8	8.6
1981	447.0	438.2	−8.8
1982	455.7	476.6	20.9
1983	507.6[b]	526.1	18.5
1984	557.7	553.5	−4.2
1985	598.6[b]	620.3	21.7

Source: Annual Report of the Board of Governors of the Federal Reserve System, various issues.

[a]Midpoints of target ranges are calculated using the fourth quarter to fourth-quarter target growth ranges except where noted. Figures refer to $M1$-A prior to 1980.

[b]Reflects change in target growth range in July.

announced new growth targets using the actual, above target, second-quarter value of $M1$ as the new base (Paul Volcker, 1985).

The practice of automatically allowing base drift to occur has long been attacked by monetarists.[4] They have frequently criticized base drift as a major impediment to any consistent policy of price stabilization in the longer run, and general economic stability in the shorter run. By letting "bygones be bygones,"[5] base drift, it is argued, essentially ratifies short-run deviations from target and hinders the achievement of stable money growth and prices over longer periods. The 1985 *Economic Report of the President* brought renewed attention to these criticisms.[6] The *Report* proposed that the Federal Reserve base its current year's target ranges on the midpoint of the previous year's target range for the fourth quarter. This alternative procedure, first suggested by William Poole (1976), would result in the complete elimination of base drift.

The purpose of the present paper is to examine the factors which determine optimal base drift. It will be shown that base drift is not necessarily inconsistent with a policy committed to price stability. In fact, such a policy objective is likely to require at least some degree of base drift. Only in special cases will either zero base drift, as advocated by the Council of Economic Advisers, or full base drift, as practiced by the FOMC, be consistent with price stability. The reason for this is that permanent shifts in the quantity of money demanded, due either to velocity shocks or income disturbances, require the real quantity of money to adjust to the new level of real money demand. If the nominal quantity of money is not adjusted, the price level must respond to equilibrate money demand and supply. Since U.S. data suggest that both real income and velocity are subject to permanent shocks,[7] price level stability requires that the nominal stock of money not be held to a fixed path. Deviations of money from target are noisy signals of these permanent shocks, and such deviations should be partially accommodated by adjusting the target path.

In Section I, a simple macro model is used to illustrate the factors which influence optimal drift. The final section summarizes the paper's basic points and uses the model's

[4]William Poole (1976) was one of the earliest to criticize automatic base drift. A more recent call for an end to base drift is found in the report of the Shadow Open Market Committee (1985). An extended criticism of base drift can be found in Alfred Broaddus and Marvin Goodfriend (1984). Milton Friedman (1985) has likened the Fed's practice to that of a marksman who always hits the bull's-eye by painting the target after taking his shot. See also Friedman (1982, p. 109; 1984, p. 37) and Bennett McCallum (1984, p. 123).

[5]Stephen Axilrod, Staff Director for Monetary and Financial Policy of the Board of Governors of the Federal Reserve System, has defended base drift by claiming "It is where you start that is important" (1982, p. 141).

[6]See, for example, the *New York Times*, March 4, 1985. For recent attacks on base drift in the popular press, see Poole (1985) and Friedman (1985).

[7]See Charles Nelson and Charles Plosser (1982) who show, using annual data, that real *GNP* and velocity are subject to permanent shocks. Velocity is also found to be a random walk by J. P. Gould et al. (1978). J. C. Kim (1985) finds using annual data that velocity in the U.K. is also a random walk.

results to critique the FOMC's current policy of complete base drift.

I. Optimal Base Drift

Let y_t, v_t, p_t, and m_t denote the natural logs of real output, velocity, the price level, and the nominal money stock, respectively. The evolution of income, velocity, and the price level is assumed to be given by equations (1)–(4):

$$(1)\quad y_t = \bar{y}_t + \alpha(p_t - {}_{t-1}p_t) + \beta(p_t - {}_{t-2}p_t) + \varepsilon_t,$$

$$(2)\quad (1-B)\bar{y}_t = u_t,$$

$$(3)\quad y_t = m_t - p_t + v_t,$$

$$(4)\quad (1-B)v_t = \phi_t + \psi_t - \psi_{t-1},$$

where B is the lag operator, ${}_sp_t$ is the minimum variance, linear forecast of p_t formed at time s, and ε, u, ϕ, and ψ are mean zero, identically and independently distributed disturbance terms. For simplicity, the disturbances are assumed to be independently distributed. The information set upon which expectations formed at time t are based consists of all current and lagged values of y, p, v, and m.

Equation (1) is an aggregate supply function in which output is a positive function of price expectational errors. Stanley Fischer (1977) shows how such an aggregate supply equation would arise in the presence of multiperiod (in this case two periods) overlapping contracts. This formulation insures that the monetary authority's choice of a target for the nominal money supply will have real effects even when announced at the start of each period, as long as the target depends in part on information unavailable one period earlier. Equation (1) also assumes that income can be written as the sum of a permanent component, $\bar{y}_t$, and a transitory component, $\alpha(p_t - {}_{t-1}p_t) + \beta(p_t - {}_{t-2}p_t) + \varepsilon_t$. The evolution of the permanent component is governed by equation (2) which specifies that $\bar{y}$ is a random walk. Together, equations (1) and (2) reflect the evidence presented by Charles Nelson and Charles Plosser (1982) that real output is a difference stationary process. The disturbance term u_t equals the innovation in the permanent component of output. From (1), it is clear that individuals (and the monetary authority) can observe the sum $\bar{y}_t + \varepsilon_t$, but not ε_t and u_t separately.[8]

The aggregate demand side of the model is represented in (3) by a simple quantity theory equation. Since J. P. Gould et al. (1978), Nelson and Plosser, and J. C. Kim (1985) all present evidence that suggests velocity is, like real income, a difference stationary process, equation (4) assumes that v_t is the sum of a permanent component ($\bar{v}_t$) with innovation ϕ_t and a transitory component with innovation ψ_t. As with the income disturbances, neither ϕ_t nor ψ_t is individually observable, although $\bar{v}_t + \psi_t = v_t$ is.

The monetary authority is assumed to set a target for the nominal money supply at the start of each period. The target will be denoted γ_t. Since γ_t is announced at the beginning of the period, it is contained in all information sets dated $t-1$ or later. Between the dates at which the targets are announced, the policy rule used to determine the nominal supply of money is taken to be of the form

$$(5)\qquad m_t = \gamma_t + \gamma[y_t - {}_{t-1}y_t] + \xi_t,$$

where ξ_t is a white-noise control error. At the start of the period, equation (5) implies that the expected value of the money supply is γ_t. If income diverges from the monetary authority's forecast, m_t is adjusted according to the parameter γ.[9] In this framework then, deviations from target are due to a

[8] The business cycle implications of an inability to distinguish between permanent and transitory shocks are explored in Karl Brunner et al. (1980).

[9] If aggregate demand were assumed to depend on the real rate of interest, and money demand on the nominal rate, the nominal rate could be introduced into equation (5) to yield what might be a more familiar looking money supply, or policy reaction, function. Unanticipated velocity movements could be made an argument in the policy reaction function without affecting any of the paper's basic results. For simplicity, equation (5) is used.

pure control error and the current realizations of the random shocks which affect income. The policy rule reflects the assumption that agents are able to contemporaneously observe income. A policy rule similar to (5) could also arise as a result of automatic adjustments to bank borrowing from the discount window, even if the monetary authority could not observe the contemporaneous value of income.

At the start of period t, the targeted value for the nominal money supply is γ_t. The actual realized value is m_t. A policy of zero base drift would use γ_t as the base for the target path in period $t+1$, while a policy of complete base drift would use m_t as the base from which the targeted value for $t+1$ is calculated. Thus, if the target growth rate is assumed, for simplicity, to be zero, we can write $\gamma_{t+1} = \delta m_t + (1-\delta)\gamma_t + x_t$, or $\gamma_{t+1} - \gamma_t = \delta(m_t - \gamma_t) + x_t$, where δ is an index of the degree of base drift, and x_t represents any additional adjustment to the target that is independent of the past deviation $m_t - \gamma_t$. An absence of drift corresponds to $\delta = 0$, and complete drift corresponds to $\delta = 1$.

The parameter γ of the policy rule (5) and the time path of γ_t are assumed to be chosen to minimize the loss function

$$(6) \qquad L(\gamma_t, \gamma) = \left[\sigma_y^2 + \rho\sigma_p^2\right],$$

where $\sigma_y^2 = E[y_t - \bar{y}_t]^2$, and $\sigma_p^2 = E[p_t - p^*]^2$. Thus, the monetary authority attempts to minimize transitory income fluctuations around the unobserved permanent component of income, and to minimize price level fluctuations around a target price level p^*.[10] The parameter ρ captures the weight placed by the monetary authority on price stability relative to output stability.

[10]Since any constant, anticipated rate of inflation has no real effects in this model, the target path for the price level could incorporate a constant rate of change. While most of the criticisms of base drift have emphasized price stability as a policy goal, the basic conclusions of this section would continue to hold if the policymaker's objective function included inflation stability. For an argument that the monetary authority should keep the price level, not the rate of inflation, on target, see Robert Hall (1984).

In order to evaluate the monetary authority's loss function, the model is first solved for the rational expectations equilibrium price function for arbitrary values of the parameters in the policy rule. The optimal values of the policy parameters γ_t and γ are then derived by minimizing the loss function.

Using (1) and (5) to eliminate y_t and m_t from (3), and noting that (1) implies ${}_{t-1}y_t = {}_{t-1}\bar{y}_t + \beta({}_{t-1}p_t - {}_{t-2}p_t)$, yields an expression for the equilibrium price level:

$$(7) \quad p_t = [1+(\alpha+\beta)(1-\gamma)]^{-1} \times[\gamma_t - (1-\gamma)(\bar{y}_t + \varepsilon_t) - \gamma_{t-1}\bar{y}_t + (\bar{v}_t + \psi_t) + (\alpha(1-\gamma) - \beta\gamma)\,{}_{t-1}p_t + \beta\,{}_{t-2}p_t + \xi_t].$$

Equation (7) can be solved, using the method of undetermined coefficients, to yield the rational expectations equilibrium price function:

$$(8) \quad p_t = k[\bar{v}_t - {}_{t-1}\bar{v}_t + \psi_t + \xi_t] - (1-\gamma)k[\bar{y}_t - {}_{t-1}\bar{y}_t + \varepsilon_t] + [1/(1+\beta)][{}_{t-1}\bar{v}_t - {}_{t-1}\bar{y}_t + \gamma_t] + [\beta/(1+\beta)][{}_{t-2}\bar{v}_t - {}_{t-2}\bar{y}_t + {}_{t-2}\gamma_t],$$

where $k = [1+(\alpha+\beta)(1-\gamma)]^{-1}$. It will be assumed that $k > 0$. From (1) and (8), the equilibrium level of output is given by

$$(9) \quad y_t = \bar{y}_t + (\alpha+\beta)k[\bar{v}_t - {}_{t-1}\bar{v}_t + \psi_t + \xi_t] - (\alpha+\beta)(1-\gamma)[\bar{y}_t - {}_{t-1}\bar{y}_t + \varepsilon_t] + \varepsilon_t + [\beta/(1+\beta)]\left[({}_{t-1}\bar{v}_t - {}_{t-2}\bar{v}_t) - ({}_{t-1}\bar{y}_t - {}_{t-2}\bar{y}_t) + (\gamma_t - {}_{t-2}\gamma_t)\right].$$

From the monetary authority's loss function, (6), and an inspection of the terms in (8) and (9) involving γ_t or expectations of γ_t, it is clear that the optimal choice for the

target value of the nominal money stock is

$$(10)\qquad \gamma_t = {}_{t-1}\bar{y}_t - {}_{t-1}\bar{v}_t + p^*.$$

The target level of m_t should be adjusted to reflect current estimates of permanent real output and permanent velocity.

It will be useful to postpone temporarily the consideration of the optimal choice of γ and just assume, as seems to be the actual case in the United States, that the monetary authority allows the money supply to vary procyclically; that is, $\gamma > 0$. To derive the optimal degree of base drift for arbitrary γ, note first that from (10),

$$(11)\qquad \gamma_{t+1} - \gamma_t = [{}_t\bar{y}_{t+1} - {}_{t-1}\bar{y}_t] - [{}_t\bar{v}_{t+1} - {}_{t-1}\bar{v}_t].$$

J. F. Muth (1960) shows that for the assumed stochastic processes generating y and v,

$$(12a)\qquad {}_{t-1}\bar{y}_t = \lambda_y \sum_{i=0}^{\infty} (1-\lambda_y)^i \times [\bar{y}_{t-1-i} + \varepsilon_{t-1-i}]$$

$$(12b)\qquad {}_{t-1}\bar{v}_t = \lambda_v \sum_{i=0}^{\infty} (1-\lambda_v)^i v_{t-1-i}$$

where

$$\lambda_y = [1 + \sigma_u^2/4\sigma_\varepsilon^2]^{1/2}(\sigma_u/\sigma_\varepsilon) - (\sigma_u^2/2\sigma_\varepsilon^2),$$

$$\lambda_v = [1 + \sigma_\phi^2/4\sigma_\psi^2]^{1/2}(\sigma_\phi/\sigma_\psi) - (\sigma_\phi^2/2\sigma_\psi^2),$$

and σ_x^2 denotes the variance of x. Both λ_y and λ_v are positive and increasing in the ratio of the variance of the permanent innovation to the variance of the transitory innovation to $\bar{y}+\varepsilon$ and v, respectively. Equations (12a) and (12b) imply that ${}_t\bar{y}_{t+1} - {}_{t-1}\bar{y}_t = \lambda_y(\bar{y}_t + \varepsilon_t - {}_{t-1}\bar{y}_t)$ and ${}_t\bar{v}_{t+1} - {}_{t-1}\bar{v}_t = \lambda_v(v_t - {}_{t-1}\bar{v}_t)$. Using these results, (11) becomes

$$(11')\qquad \gamma_{t+1} - \gamma_t = \lambda_y(\bar{y}_t + \varepsilon_t - {}_{t-1}\bar{y}_t) - \lambda_v(v_t - {}_{t-1}\bar{v}_t).$$

Equation (11′) is the fundamental relationship governing the appropriate adjustment in the target base. In order to determine the optimal degree of base drift, it is necessary to relate this adjustment in the target to the deviation of the money supply from target in period t. Equation (5) and (11′) can be used to write

$$(13)\qquad \gamma_{t+1} - \gamma_t = (\lambda_y/\gamma k)(m_t - \gamma_t) - (\lambda_v + \lambda_y(\alpha+\beta))(\bar{v}_t + \psi_t - {}_{t-1}\bar{v}_t) - (\lambda_y/\gamma)(1+\alpha+\beta)\xi_t.$$

Equation (13) gives the optimal base adjustment as a function of the deviation from target during the period just ending, $m_t - \gamma_t$, the velocity surprise in period t, and the monetary control error. To interpret this equation, it is helpful to start with a special case. Suppose velocity is constant and there are no control errors. In this case, (13) becomes

$$(14)\qquad \gamma_{t+1} - \gamma_t = (\lambda_y/\gamma k)(m_t - \gamma_t),$$

so that the optimal degree of base drift is $\delta = \lambda_y/\gamma k > 0$. The index δ depends on both the structural parameters of the model (α and β appear in k), the monetary authority's behavior between target announcements (via γ), and, through λ_y, the relative variances of the permanent and transitory shocks to real output.

In standard aggregate models, it is common to assume that output follows a stationary process. In the present model, this would amount to assuming $\sigma_u^2 = 0$. This, in turn, implies that $\lambda_y = 0$. Thus, if income is not subject to any permanent shocks, optimal base drift is zero.[11] As mentioned previously, however, recent work by Nelson and Plosser suggests that output is difference stationary. In fact, they estimate the ratio of the

[11] If $\gamma = (1+\alpha+\beta)/(\alpha+\beta)$, then $\delta = 0$ even if $\lambda_y > 0$. Since $\alpha+\beta$ is usually estimated to be fairly small, this would require a large value for γ. Economically, this special case seems of little interest.

variance of permanent shocks to income to the variance of transitory shocks to be quite large. In order to keep the average price level constant, the monetary authority must let the nominal stock of money vary with movements in the real demand for money arising from permanent shocks to real income. If the actual money stock differs from its targeted level, the deviation is likely to be partially due to such a permanent shock, calling for an adjustment in the target for the nominal money supply. If base drift is never allowed, the price level must adjust to insure that the real quantity of money equals the real demand for money.

As (14) shows, δ is increasing in λ_y and decreasing in γ. With a larger λ_y, more of any income surprise is treated as due to a permanent shock, calling for more drift. A larger γ implies that m responds more within the period to output shocks; consequently, a smaller adjustment in the base is required.

Equation (14) holds only in the absence of control errors and with a constant velocity. When these conditions do not hold, equation (13) can be written

$$(15)\quad \gamma_{t+1}-\gamma_t=\delta(m_t'-\gamma_t) - \lambda_v(\bar{v}_t+\psi_t-{}_{t-1}\bar{v}_t),$$

where the optimal degree of base drift, δ, is exactly the same as in (14), but drift is now defined relative to a measure of "shift-adjusted" money:

$$(16)\quad m_t'=m_t-\xi_t-\gamma k(\alpha+\beta)\times(\bar{v}_t+\psi_t-{}_{t-1}\bar{v}_t+\xi_t).$$

According to equation (16), two types of adjustments need to be made to the realized money stock before determining optimal drift. The first is a one-for-one adjustment for any control errors. This shows clearly that in calculating base drift, bygones are not bygones: control errors should be fully offset. The second adjustment, given by the last term in (16), arises because of the induced response of the money supply to unanticipated income movements through the policy rule (5). Velocity shocks and control errors affect the excess supply of money and thereby affect income. This, through the policy reaction function, has an automatic impact on m_t. As (16) shows, m_t should be adjusted to offset changes arising from this source before drift is determined. Because of the induced impact via income, the total adjustment for control errors is more than one-for-one.

In addition to shift adjusting the realized money supply, equation (15) indicates that velocity shocks require a base change even if shift-adjusted m_t equals the target γ_t. The last term in (15) is equal to the fraction of any velocity shock that is expected to be permanent. The target for the money supply should be fully adjusted to reflect estimated permanent shifts in velocity. If all velocity shocks are transitory ($\lambda_v=0$), this term is zero. Transitory velocity shocks still have an effect on target levels, though, through the shift adjustment in equation (16). At the other extreme, if velocity is a random walk ($\lambda_v=1$), the monetary authority should fully offset changes in velocity, plus shift-adjust the realized money supply for velocity shocks.

So far, the results have all been predicated on an arbitrary value for γ. However, the monetary authority's choice of γ affects the short-run tradeoff between output stabilization and price stabilization and, hence, the value of the monetary authority's loss function given in (6). An optimal setting for γ can be found by minimizing $[\sigma_y^2+\rho\sigma_p^2]$ with respect to γ. As long as the target money supply is chosen according to (11), equations (8) and (9) imply $p_t-p^*=z_t$ and $y_t-\bar{y}_t=(\alpha+\beta)z_t+\varepsilon_t$, where $z_t=k[\bar{v}_t-{}_{t-1}\bar{v}_t+\psi_t+\xi_t]-(1-\gamma)k[\bar{y}_t-{}_{t-1}\bar{y}_t+\varepsilon_t]$. Hence, the loss function is independent of the base parameter γ_t. Setting the partial derivative of the loss function with respect to γ equal to zero and solving for the optimal policy response γ^*, one finds that

$$(17)\quad \gamma^*=1-(\alpha+\beta)\times\left[\Omega\left(\sigma_v^2+\sigma_\psi^2+\sigma_\xi^2\right)+\sigma_\varepsilon^2\right]/\left[\Omega\sigma_y^2+\rho\sigma_\varepsilon^2\right],$$

where $\Omega=(\alpha+\beta)^2+\rho$,

and
$$\sigma_{\bar{v}}^2 = E[\bar{v}_t - {}_{t-1}\bar{v}_t]^2 = \left[\sigma_\phi^2 + \lambda_v^2\sigma_\psi^2\right]/[\lambda_v(2-\lambda_v)],$$
$$\sigma_{\bar{y}}^2 = E[\bar{y}_t - {}_{t-1}\bar{y}_t]^2 = \left[\sigma_u^2 + \lambda_y\sigma_\varepsilon^2\right]/\lambda_y^2(2-\lambda_y)]$$

are the variances of the prediction errors of permanent velocity and income. It is useful to recall that the covariances among all the underlying shocks have been assumed to equal zero.

If the aggregate demand disturbances (velocity shocks and monetary control errors) are large relative to supply shocks, income and prices will tend to move together, and the optimal γ is likely to be negative as the monetary authority works to offset demand shocks. In this case, the formula for δ shows that drift should be negative. If a supply shock causes a rise in income, m_t will undershoot its target, and the appropriate response by the monetary authority is to increase its target for the money supply. On the other hand, γ^* increases with the variance of permanent supply innovations as positive shocks to output require an increase in aggregate demand to prevent the price level from changing. The impact of σ_ε^2, the variance of transitory supply shocks, on γ^* is ambiguous and depends on the weight the monetary authority places on price stabilization. If ρ is large, γ^* will increase with σ_ε^2.

Using equation (17), it is possible to derive the appropriate degree of base drift when γ is optimally chosen. Calling this δ^*, some simple algebra yields

$$(18)\quad \delta^* = \lambda_y\Omega\left[\sigma_{\bar{y}}^2 + \sigma_\varepsilon^2 + (\alpha+\beta)\left(\sigma_{\bar{v}}^2 + \sigma_\psi^2 + \sigma_\xi^2\right)\right] \Big/ \left[\Omega\sigma_{\bar{y}}^2 + (\rho-\alpha-\beta)\sigma_\varepsilon^2 - (\alpha+\beta)\Omega\left(\sigma_{\bar{v}}^2 + \sigma_\psi^2 + \sigma_\xi^2\right)\right].$$

In general, δ^* depends on the error variances, the structural parameters, and the monetary authority's preferences for price versus output stability as measured by ρ. A sufficient condition for $\delta^* = 0$ continues to be $\lambda_y = 0$ (i.e., no permanent shocks to supply). A sufficient condition for $\delta^* = 1$ is $\lambda_y = 1$ (no transitory supply shocks) *and* no velocity disturbances or control errors. Also, even though $\delta = 0$ when $\lambda_y = 0$, this only implies that the target for $t+1$ should not depend on deviations of m_t from target; it does not imply the target base should remain constant. This latter condition, as (11′) shows, requires that both λ_y and λ_v equal zero. While the expression for δ is a complicated function of all the parameters of the model, it is important to recall that the fundamental equation determining the optimal base change is (11′) which depends only on λ_y and λ_v. Because income movements affect the contemporaneous deviation of money from target, while the policy rule affects income movements, much more information is needed, as (18) shows, if the base change is to be related to the target deviation.

II. Interpretation

The results of Section I were derived within the context of a simplified aggregate model, but the basic conclusions suggested by the model will hold more generally. To the extent that deviations of the nominal money stock from its target path reflect permanent shocks to the economy, a policy of price stability requires some degree of base drift. While zero drift will not be optimal, neither will complete base drift generally be consistent with a policy of price stability. It is interesting, though, to consider how the actual practice of the FOMC compares to the implications derived for optimal drift.

Equations (15) and (16) show that the appropriate adjustment in the target for the nominal money supply at the start of each period depends very much on the reason there may have been a divergence between actual money and targeted money. In the present model, supply shocks, velocity shifts, and control errors can all cause m_t to differ from γ_t, and each type of disturbance calls for a different adjustment in the target base. For example, if the deviation is due to a

velocity shock, two adjustments should be made. First, there should be an offsetting movement in the base equal to the fraction of the velocity shock estimated to be permanent. This, in fact, is done indirectly by the FOMC. The Fed has in several years focused on $M1$ adjusted for velocity shifts due to financial deregulation. Such shifts occurred, for example, with the introduction of automatic transfer from savings (ATS) accounts in 1979 and the national authorization of negotiable order of withdrawal (NOW) accounts in 1981.[12] Since the FOMC has practiced complete base drift with respect to deviations of this adjusted $M1$ from target, it has essentially made what equation (16) shows to be the appropriate base change (i.e., full adjustment for estimated permanent shifts in velocity) if it has correctly estimated the permanent velocity shifts. For example, for 1981, Table 1 shows that base drift was $-\$8.8$ billion. Financial deregulation (the national authorization of NOW accounts) produced what the Fed estimated to be a permanent upward shift in velocity. Some negative drift was therefore appropriate in setting the target path for 1982. These adjustments for permanent shifts in velocity should be made even if the nominal money supply were to equal its target value for the fourth quarter.

Second, any velocity shock, whether permanent or transitory, produces an effect on the nominal money supply through the monetary authority's reaction function. This effect on the money supply must be offset if the price level in future periods is to be kept constant.

If output for period t is underestimated due to either a permanent (u_t) or a transitory (ε_t) supply shock, $\dot{m}_t$ will exceed its target γ_t. Because some fraction of this miss should be attributed to a positive realization of the permanent shock, an upward revision in the base is required for price and output stabilization.[13] Some drift is optimal in this case. However, only if all income shocks are permanent and there are no velocity shocks or control errors would complete base drift, as practiced by the Fed, be optimal.

If the money supply target is missed because of a control error, the realized money supply should be adjusted, in calculating the appropriate change in the base, in order to offset the impact of the control error. Under current Fed practice, a positive control error leaves the nominal money supply permanently higher.[14] For the model studied here, the optimal response calls for a shift adjustment that more than offsets the control error.

In its critique of base drift, the Council of Economic Advisers attributed deviations of $M1$ from target, and hence the subsequent base drift, to the Fed's attempts to smooth interest rates. As the analysis of optimal base drift shows, deviations from target due to any response of money to current shocks should not lead to a change in the base unless a permanent innovation to income or velocity is estimated to have occurred. The recent deemphasis of $M1$ by the FOMC may mean that future deviations of $M1$ will be more responsive to transitory disturbances and subject to larger control errors. This would suggest the optimal degree of base drift will decline.

While the analysis suggests that a policy of zero base drift will not be optimal, it also suggests that the current Federal Reserve practice of complete base drift is also not generally optimal. This is particularly true with respect to monetary control errors which produce deviations of the money supply from its target path. However, during periods of technological change in both the real and financial sectors, there is no a priori reason to argue that base drift is inconsistent with a policy objective of price stability. In fact,

[12] These adjustments are discussed in Barbara Bennett (1982) and Broaddus and Goodfriend.

[13] This discussion assumes $\gamma > 0$. If $\gamma < 0$, then a positive income disturbance will cause m_t to fall below γ_t. With $\delta = \lambda_y/\gamma k < 0$, this implies that γ_{t+1} should be increased. As (11′) shows, the appropriate change in the target is independent of the sign of γ, even though the sign of δ is not.

[14] Poole calls base drift "the insidious practice of sweeping monetary control errors under the rug..." (1985, p. 4).

empirical evidence suggesting the presence of permanent real and financial shocks indicates that the complete elimination of base drift is likely to be incompatible with the achievement of price stability.

REFERENCES

Axilrod, Stephen H., "Money, Credit, and Banking Debate," *Journal of Money, Credit, and Banking*, February 1982, *14*, 119–47.

Bennett, Barbara A., "'Shift Adjustments' to the Monetary Aggregates," Federal Reserve Bank of San Francisco *Economic Review*, Spring 1982, 6–18.

Broaddus, Alfred, and Goodfriend, Marvin, "Base Drift and the Longer Run Growth of *M*1: Experience from a Decade of Monetary Targeting," Federal Reserve Bank of Richmond *Economic Review*, November/December 1984, *70*, 3–14.

Brunner, Karl, "The Politics of Myopia and Its Ideology," in SOMC *Policy Statement and Position Papers*, Center for Research in Government Policy and Business, Graduate School of Management, University of Rochester, September 1983.

______, **Cukierman, Alex and Meltzer, Allan H.**, "Stagflation, Persistent Unemployment, and the Permanence of Economic Shocks," *Journal of Monetary Economics*, October 1980, *6*, 467–92.

Fischer, Stanley, "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *Journal of Political Economy*, February 1977, *85*, 191–205.

Friedman, Milton, "Monetary Policy: Theory and Practice," *Journal of Money, Credit, and Banking*, February 1982, *14*, 98–118.

______, "Monetary Policies for the 1980s," in J. H. Moore, ed., *To Promote Prosperity*, Stanford: Hoover Institution Press, 1984.

______, "The Fed Hasn't Changed Its Ways," *Wall Street Journal*, August 20, 1985.

Gould, J. P. et al. "The Stochastic Properties of Velocity and the Quantity Theory of Money," *Journal of Monetary Economics*, April 1978, *4*, 229–48.

Hafer, R. W., "The FOMC in 1983–84: Setting Policy in an Uncertain World," Federal Reserve Bank of St. Louis *Economic Review*, April 1985, 15–37.

Hall, Robert E., "Monetary Strategy with an Elastic Price Standard," in *Price Stability and Public Policy*, Kansas City: Federal Reserve Bank, 1984.

Kim, J. C., "Random Walk and the Velocity of Money: Some Evidence from Annual and Quarterly Data," *Economic Letters*, 1985, *18*, 187–90.

McCallum, Bennett T., "Credibility and Monetary Policy," in *Price Stability and Public Policy*, Kansas City: Federal Reserve Bank, 1984.

Muth, J. F., "Optimal Properties of Exponentially Weighted Forecasts," *Journal of the American Statistical Association*, June 1960, *55*, 299–306.

Nelson, Charles R. and Plosser, Charles I., "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications," *Journal of Monetary Economics*, September 1982, *10*, 139–62.

Poole, William, "Interpreting the Fed's Monetary Targets," *Brookings Papers on Economic Activity* 1:1976, 247–59.

______, "Monetary Targeting is Dead; Another Boom-Bust Cycle is Next," *American Banker*, August 16, 1985.

Volcker, Paul A., "Monetary Policy Objectives for 1985," Testimony before Congress, July 17, 1985.

Board of Governors of the Federal Reserve System, *Annual Report*, various issues.

Shadow Open Market Committee (SOMC), "Policy Statement," Center for Research in Government Policy and Business, Graduate School of Management, University of Rochester, March 1985.

U.S. Council of Economic Advisors, *Economic Report of the President*, Washington, 1985.

Prizes and Incentives in Elimination Tournaments

By SHERWIN ROSEN*

Contestants who succeed in attaining high ranks in elimination career ladders rest on their laurels in attempting to climb higher, unless top-ranking prizes are given a disproportionate weight in the purse. A large first-place prize gives survivors something to shoot for, independent of past performances and accomplishments.

Several recent papers have clarified the problem of incentives when competitors are paid on the basis of rank or relative performance (Edward Lazear and myself, 1981; Jerry Green and Nancy Stokey; 1983; Barry Nalebuff and Joseph Stiglitz, 1984; Bengt Holmstrom, 1982; James Malcomson, 1984; Lorne Carmichael, 1983; Mary O'Keefe, W. Kip Viscusi, and Richard Zeckhauser, 1984). The main focus so far has been to examine the economic efficiency of these schemes. However, a much longer tradition in statistics views relative comparisons as an experimental design for ranking and selecting contestants. These two views are joined in this work.

I investigate the incentive properties of prizes in sequential elimination events, where rewards are increasing in survival. The inherent logic of these experiments is to determine the best contestants and promote survival of the fittest; and to maintain the "quality of play" as the game proceeds through its stages. Athletic tournaments immediately come to mind, but much broader interest in this class of problems arises from its potential application to career games, where the tournament analogy is supported (James Rosenbaum, 1984). Many organizations have a triangular structure (for example, Martin Beckmann, 1978) and most top level managers come up through the ranks (Kevin J. Murphy, 1984). A career trajectory is, in part, the outcome of competition among peers to attain higher ranking and more remunerative positions over the life cycle. The structure of rewards influences the nature and quality of competition at each stage of the game.

What needs to be explained is the marked concentration of rewards in the top ranks. For example, the top four ranks receive 50 percent or more of the total purse in tennis tournaments. Concentration is less extreme in the executive labor market, but nonetheless, earnings rise more than proportionate to rank in most firms. I show below that an elimination design requires an extra reward for the overall winner to maintain performance incentives throughout the game.

The economics of this result derives from the survival aspects of the game. A competitor's performance incentives at any stage are set by an option value. The loser's prize is guaranteed at that stage, but winning gives the option to continue on to all successive rungs in the ladder. There are fewer steps remaining to be attained as the game proceeds, and the option value plays out. It expires in the final match because advancement opportunities vanish. At that point, the difference in prize money between winning and losing must incorporate the equivalent of the survival option that maintained incentives at earlier stages. The extra weight of rewards at the top is due to the no-tomorrow aspects of the final stage of the game. It extends the horizon of players surviving to those stages, and makes the game appear of

*Department of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637. I am especially indebted to Barry Nalebuff for many suggestions that greatly improved this work, to Edward Lazear, Kevin M. Murphy, David Pierce, and Nancy Stokey for advice on a number of points, and to Gary Becker, James Friedman, Sandy Grossman, and John Riley for comments on an initial draft. Robert Tamura was my research assistant. This project was supported by the National Science Foundation.

infinite length to a contestant, *as if* there are always many steps left to attain, no matter how far one has climbed in the past. This result obviously bears a family resemblance to the role of a "pension" in a finitely repeated principal and agent problem (Gary Becker and George Stigler, 1984; Lazear, 1981).

Section I describes the game, and Section II sets forth the nature of contestants' strategies. Sections III and IV analyze the problem when the inherent talents of competitors are known, while Section V analyzes the case where talents are unknown.

I. Design of the Game

For analytical tractability and simplicity, the ideas are best revealed by a paired-comparison structure, as in a tennis-ladder. The tournament begins with 2^N players and proceeds sequentially through N stages. Each stage is a set of pairwise matches. Winners survive to the next round, where another pairing is drawn, and losers are eliminated from subsequent play. Half are cut from further consideration at each stage. Thus, in a career game, those eligible for promotion to some rank have attained the rank immediately below it. Those who are passed over at any stage are out of the running for further promotions. The top prize W_1 is awarded to the winner of the final match, who has won N matches overall. The loser of the final match achieves second place overall and is awarded prize W_2 for having won $N-1$ matches. Losers of the semifinals are both awarded W_3, etc.

Define s as the number of stages remaining to be played. Then all players eliminated in a match where s stages remain are awarded prize W_{s+1}. Define the interrank spread $\Delta W_s = W_s - W_{s+1}$ as the marginal reward for advancing one place in the final ranking. These increments determine incentives to advance through the stages. Prizes are increasing in survival: $\Delta W_s > 0$ for all s.

I am concerned in this work with studying how prizes affect performance and selection, and with finding some characterizations of the relative reward structure required to maintain incentives as the game proceeds. This is a piece of a larger problem of the "optimal" prize structure, the study of which requires specifying how incentives affect the social value of the game. These complex matters are not well understood. So rather than tying results to an arbitrary input-output technology, a common feature of the larger problem obviously requires that players work at least as hard, if not harder, in the later stages of the game as in the early stages. We don't want contestants to lay down near the end. For example, in a hierarchical organization, the decisions made at the top are more important than those made further down the pyramid (see my 1982 paper): shirking and lack of talent have more serious consequences at the top of the organization than at the bottom.

Rank-order schemes are encountered when individual output and input are difficult to measure on a cardinal scale, an inherent feature of managerial and many other types of talent; or when common background noise contaminates precise individual assessments of value-added. Competition is inherently head-to-head in most athletic games, and cardinality in any sense other than probability of winning has little meaning. Ordinality is inherent because the point scores used to calibrate performance contain many arbitrary elements, as in a classroom test. Many of these same considerations apply to selection of managerial talents though competition is not strictly paired comparisons.

Given the rules, these issues may be finessed for studying the connection between prizes, incentives, and selection by specifying how players' actions affect the probability of winning. Let i index a player and let j index an opponent in some match. Consider a game in which there are m types of players. Index the ability type of the i player by I and the ability type of the opponent j player by J: Both I and J take on m possible values, $1,2,\dots,m$, with $m \leq 2^N$. Let x_{si} and x_{sj} denote the intensity of effort expended by players i and j in a match when s stages remain to be played, and let γ_I and γ_J represent their abilities or natural talents for the game. Then $P_s(I,J)$ is the probability that a player of type I wins in a match against a player of type J (possibly the same

type) with

$$(1)\quad P_s(I,J)=\frac{\gamma_I h(x_{si})}{\gamma_I h(x_{si})+\gamma_J h(x_{sj})},$$

where $h(x)$ is increasing in x and $h(0)\geq 0$. A player increases the probability of winning the match by exerting greater effort, given the talent and effort of the opponent and own talent. To simplify the problem, the win technology is assumed identical at every stage (s enters only through the x's).

When both players exert the same level of effort, the win probability is $P_s(I,J)=\gamma_I/(\gamma_I+\gamma_J)$, and its inverse is a bookmaker's "morning line" or "true-to-form" actuarially fair payoffs per dollar bet on player-type I. Notice from (1) that common, multiplicative environmental factors do not affect $P_s(I,J)$. Let the common factor multiply $\gamma h(x)$ for both players. Then whether the commonality is match-, stage-, or tournament-specific, it factors out of the probability calculation and has no effect on either incentives or selection. Equation (1) is a logit when $h(x)$ is exponential. Alternatively, think of $\gamma h(x)$ as the arrival rate of a Poisson process. Then (1) can be given a racing game interpretation, as in the recent literature on patent races (Glenn Loury, 1979).

II. Strategies

A player's decision of how much effort to expend in any match depends on weighing the benefit of greater effort (increasing the probability of surviving) against its costs. There are two complications. First, the value of advancing depends on how the player assesses future effort should eligibility be maintained. This forward-looking effect is analyzed by backward recursion. Second, current actions depend on the anticipated behavior of the current and all future possible opponents. The sequential character of the game allows this to be analyzed by adopting Nash noncooperative strategies as the equilibrium concept. Discounting between stages is ignored and risk neutrality is assumed.

Define $V_s(I,J)$ as the value to a player of type I of playing a match against an opponent of type J when s possible stages remain to be played. Assume, for now, that all players' talents are common knowledge. Let $c(x)$ be the cost of effort in any match, assumed identical for all players, $c'(x)>0$, $c''(x)\geq 0$, and $c(0)=0$. The value of the match consists of two components: one is W_{s+1}, the prize earned if the match is lost and the player is eliminated, an event which occurs with probability $1-P_s(I,J)$. The other is the value of achieving a final rank superior to $s+1$ if the match is won. Let $EV_{s-1}(I)$ represent the expected value of eligibility in the next stage. This is a weighted average over J of $V_{s-1}(I,J)$, where the weights are the probabilities that the I player will confront an opponent of type J in the next stage. These probabilities depend on the activities of players in other matches and the rules for drawing opponents at each stage. The probability of continuation is $P_s(I,J)$, and costs $c(x)$ are incurred for either outcome, so the fundamental equation for this problem is

$$(2)\quad V_s(I,J)=\max_{x_{si}}\big[P_s(I,J)EV_{s-1}(I)+(1-P_s(I,J))W_{s+1}-c(x_{si})\big].$$

The max in (2) is understood on Nash assumptions as conditioned on the given current and expected future efforts of all other players remaining alive at s and on the optimum actions taken by the player in question in subsequent matches.

Substituting (1) into (2) and differentiating with respect to x_{si} yields the first-order condition

$$(3)\quad \gamma_I\gamma_J h_j h_i'/(\gamma_I h_i+\gamma_J h_j)^2\times\big[EV_{s-1}(I)-W_{s+1}\big]-c_i'=0,$$

where $h_i=h(x_{si})$, $h_i'=dh(x_{si})/dx_{si}$, etc. The second-order condition is

$$(4)\quad D=c_i'\big[(h_i''/h_i')-2\gamma_I h_i'/(\gamma_I h_i+\gamma_J h_j)\big]-c_i''<0.$$

Note that (4) allows $h''>0$ so long as it is bounded. There is also a global condition.

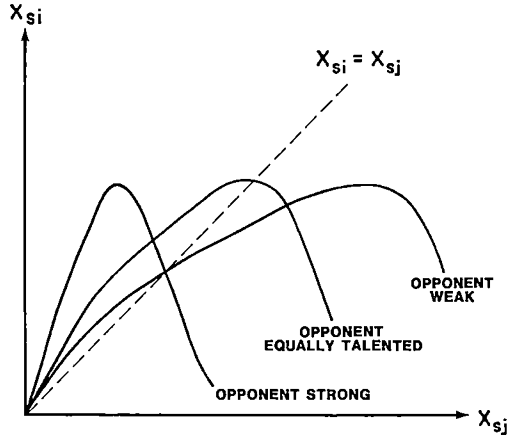


FIGURE 1

Equation (3) indicates that effort in any match is controlled by $EV_{s-1}(I) - W_{s+1}$. This difference between winning and losing must be positive for the player to have an interest in maintaining eligibility into the next stage. Otherwise, it is best to default and exert no effort.

Equation (3) defines the best-response function for player i. Differentiating with respect to the current opponent's effort,

$$(5)\quad \partial x_i/\partial x_j = \frac{(h_j'/h_j)c_i'}{-D(\gamma_I h_i + \gamma_J h_j)} \times (\gamma_I h_i - \gamma_J h_j).$$

Player i's best reply is increasing in x_j when x_j is small enough, but is decreasing when the opponent's effort is sufficiently large. It has a turning point at $\gamma_I h(x_i) = \gamma_J h(x_j)$. The turning point occurs at $x_i = x_j$ for equally talented players ($\gamma_I = \gamma_J$). It turns at some value $x_j > x_i$ when i is playing a weaker opponent ($\gamma_I > \gamma_J$) and it turns at some $x_j < x_i$ when the opponent is the stronger player ($\gamma_I < \gamma_J$). See Figure 1. Analysis is confined to pure-strategy equilibria.[1]

III. Incentive Maintaining Prizes: Equally Talented Contestants

The solution is transparent when all players are equally talented (there is only one type). Then $EV_{s-1}(I) = V_{s-1}$, because each player knows for sure that an opponent of equal skill will be confronted at every stage. From Figure 1, the best-reply function is the same for all players and has a turning point at $x_i = x_j$. Therefore, the equilibrium is symmetric: $x_{si} = x_{sj} = x_s$ for all i and j and $P_s = 1/2$ in equilibrium. Each match is a close call in expected value. The common level of effort when s stages remain which satisfies (3) is

$$(6)\quad (V_{s-1} - W_{s+1})(h'(x_s)/h(x_s))/4 = c'(x_s).$$

Define the elasticities

$$(7)\quad \eta(x) = xh'(x)/h(x),$$
$$\varepsilon(x) = xc'(x)/c(x),$$
$$\mu(x) = \eta(x)/\varepsilon(x).$$

Then (6) becomes

$$(8)\quad (V_{s-1} - W_{s+1})\mu(x_s)/4 = c(x_s).$$

Substituting (8) into (2) and using $P_s = 1/2$,

$$(9)\quad V_s = (1/2)(1 - \mu(x_s)/2) \times (V_{s-1} - W_{s+1}) + W_{s+1} = \beta_s V_{s-1} + (1 - \beta_s) W_{s+1},$$

[1]The best reply may jump down to zero at some point because either (4) fails beyond that point or default ($x_i = 0$) is a global optimum while (3) is local. Pure strategies characterize equilibrium when these jumps occur (if they do) at sufficiently large x_i. This requires certain bounds on the curvature of the $h(x)$ and $c(x)$ functions and some limits on the degree of heterogeneity (the γ's) among players (for example, a very weak player might just lay down against a very strong one). The rules of the game determine $c(x)$ and $h(x)$ (see O'Keeffe et al.). Rules and initial screening of entrants must be suitably constrained to guarantee pure-strategy equilibria. Nalebuff and Stiglitz analyze random strategies in one-shot games. Stephen Bronars (1985) shows that a weak player might employ a riskier strategy against a stronger opponent, but (1) is not suitably parameterized to consider this.

where

$$(10) \quad \beta_s = (1/2)(1-\mu(x_s)/2).$$

The recursion (9) holds if (3) is a global maximum, and no player has incentives to default from x_s defined by (6). This requires, from (9), that $V_s - W_{s+1} = \beta_s(V_{s-1} - W_{s+1}) > 0$. Otherwise taking the sure loss is a better choice. Therefore $\beta_s > 0$, or, from (10) and (7), $\eta(x_s)/2\varepsilon(x_s) < 1$, or $\mu(x) < 2$. There is no pure-strategy equilibrium in this game if any player has an incentive to default.

The sense of the no-default condition $\eta(x)/\varepsilon(x) < 2$ is related to the problem of an arms race. If the elasticity of response of effort is large relative to the elasticity of its cost, then players' efforts to win results in a negative sum game in pure strategies. It is not optimal to default if the opponent does, but at the local equilibrium the costs of contesting have been escalated so much that both want to default. In fact, (9) implies that for given prizes, players are better off when there is less scope for actions to affect outcomes: V_s is decreasing in $\mu(x)$, so the rules of the game must be devised to balance two conflicting forces: games which greatly constrain the effect of actions on outcomes are inefficient and unproductive: whereas competition is destructive if these constraints are relaxed too much.[2]

Assuming $0 < \beta_s < 1$, for all s and using $V_0 = W_1$ as a boundary condition, the solution to (9) is

$$(11) \quad V_s = (\beta_1\beta_2\ldots\beta_s)\Delta W_1 + (\beta_2\ldots\beta_s)\Delta W_2 + \ldots + \beta_s \Delta W_s + W_{s+1}.$$

The value of maintaining eligibility at any stage is the sure prize the player has guaranteed by surviving that long, plus the discounted sum of successive interrank rewards that may be achieved in future matches. Herein lies the "option" value of an elimination design. Manipulating (11) yields an expression for $V_{s-1} - W_{s+1}$, which controls performance incentives, from (8):

$$(12) \quad (V_{s-1} - W_{s+1}) = (\beta_1\ldots\beta_{s-1})\Delta W_1 + (\beta_2\ldots\beta_{s-1})\Delta W_2 + \ldots + \Delta W_s.$$

Incentives are determined by the discounted sum of interrank spreads.

What reward structure maintains incentives to perform at a common value throughout all stages of the game? Here $x_s = x^*$ for all s and $\beta_s = \beta$ is a constant for all s, from (10). Then (12) implies

$$(13) \quad (V_{s-1} - W_{s+1}) - \beta(V_{s-2} - W_s) = \Delta W_s \quad \text{for } s = 2,3,\ldots,N.$$

Constant performance requires that $(V_{s-1} - W_{s+1})$ is a constant. Suppose $V_{s-1} - W_{s+1} = k$, where k is determined so that $x_s = x^*$ solves (3). Then from (13),

$$(14) \quad k(1-\beta) = \Delta W = \Delta W_s \quad \text{for } s = 2,3,\ldots,N$$

and since final-round effort depends only on ΔW_1, from (12),

$$(15) \quad k = \Delta W_1 = \Delta W/(1-\beta) > \Delta W.$$

The incentive-maintaining prize structure requires a constant interrank spread from second place down, from (14). However, it requires a larger interrank spread at the top, from (15). Prizes rise *linearly* in increments $\Delta W = k(1-\beta)$ from rank $N+1$ up through rank 2, but the first place prize takes a distinct *jump* out of sync with the general linear pattern below it. The incentive-maintaining prize distribution weighs the top prize more heavily than the rest.[3] See Figure 2.

[2] Contestants have incentives to introduce new techniques and styles of play to create a winning edge. These are sources of technical change in career games. Athletic games use a supreme authority to maintain the integrity of the game. Innovations which escalate the collective costs of competition relative to social value are prohibited.

[3] This analysis determines only relative prizes across ranks, not their absolute level. More structure on technologies and the social value of the game must be introduced to examine the latter (for example, see Lazear's and my article). Here we require that the purse is large enough to support $V_s > 0$ for all s. This implies

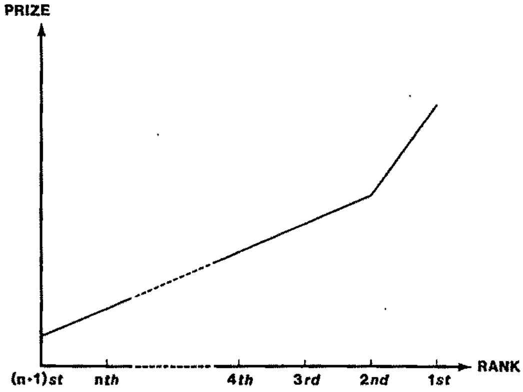


FIGURE 2

The proof supports the economic interpretation of this surprising conclusion. The final-round spread has to replace the earlier option value of achieving possible higher ranks at earlier stages. Substitute (14) and (15) into (12):

$$(16)\quad (V_{s-1}-W_{s+1})$$
$$=\beta^{s-1}\Delta W_1+\Delta W(\beta^{s-2}+\beta^{s-3}+\dots+1)$$
$$=\Delta W\left[(\beta^{s-1}/(1-\beta))+\beta^{s-2}+\beta^{s-3}+\dots+1\right]$$
$$=\Delta W(1+\beta+\beta^2+\beta^3+\dots)\quad \text{for all } s,$$

The extra increment at the top converts the value of the difference between winning and losing at each stage into a perpetuity of constant value at all stages. It effectively extends the horizon of the players and makes them behave *as if they are in a game which continues forever*. This horizon-extending feature of the top prize is one of the reasons why observed rewards are concentrated toward the top ranks. It is clear that concentrating even more of the purse on the top creates incentives for performance to increase as the game proceeds through its stages. For example, if the winner takes all, then every term other than the one in ΔW_1 in (12) vanishes and the difference in value between winning and losing increases as the game proceeds, through the force of discounting: effort is smallest in the first stage and largest in the finals.

The result in (16) and (17) is robust to a number of modifications:

(*i*) *Risk Aversion*. Suppose preferences take the additive form $U(W)-\Sigma_s c(x_s)$, where $c(x)$ is as before and $U(W)$ is increasing, but not necessarily linear in W. Then the entire analysis goes through by replacing W_s with $U(W_s)$ wherever it appears. Incentive maintenance requires a constant difference in the utility of rewards $U(W_{s+1})-U(W_{s+2})$ in all stages prior to the finals, but still requires a jump in the interrank difference in utility of winning the finals. If players are risk averse $(U''(W)<0)$, the incentive-maintaining prize structure requires strictly increasing interrank spreads, with an even larger increment between first and second place. The prize structure is everywhere convex in rank order, with greater concentration of the purse on the top prizes than when contestants are risk neutral. The spread has to be increasing to "buy off" survivor's risk aversion and maintain their interest in advancing to higher ranks.

(*ii*) *Symmetric Win-Technologies*. The derivation of (14) and (15) rests only on that property that P_s is 1/2 in equilibrium. Hence Figure 2 is independent of the specific form of (2) and holds for *any* win technology resulting in a symmetric equilibrium. Furthermore, the result extends to more than pairwise comparisons: there might be n-way comparisons at each stage. In the Poisson case, the probability of advancing becomes $h(x_i)/\Sigma^n h(x_k)$. Then $\beta_s=(1/n)(1-(n-1)\mu(x_s)/n)$, but the logic otherwise remains unchanged.

(*iii*) *Stage Effects*. The nature of competition may vary across stages. For example, in a corporate hierarchy the pass-through rate may fall at each successive rank. Similarly, μ_s may be smaller in the later stages because higher-ranking positions are more demanding than lower-ranking ones. In

an upper bound on feasible x^*. Another upper bound on x^* is implied by contestants' outside opportunities, but is ignored.

either case, β_s decreases as the game proceeds, and interrank spreads must be increasing to undo the incentive dilution effects of greater discounting of the future, which otherwise reduces the option value of continuation. These considerations increase the convexity of the rank/reward structure.[4]

IV. Heterogeneous Contestants with Known Talents

In heterogeneous populations, elimination designs promote *survival of the fittest* and progressive elimination of weaker contenders. The conditional mean ability of survivors tends to increase as the game proceeds and differences in survivors' talents are compressed relative to the initial field. This increasing homogeneity among surviving members across stages extends the incentive-maintenance result above to the limit of the last few stages of a long game. For by continuity of the best-reply functions in ability parameters, EV_{s-1} is approximately V_{s-1} among relatively homogeneous survivors in the final stages. The extra final-round incremental prize remains necessary to maintain incentives toward the end.

This section shows that the value of the continuation option is increasing in ability, which is why the design encourages survival of the fittest. However, analysis is complicated by progressive increasing strength-of-field effects. That stronger opponents are likely to be encountered in later stages reduces the value of continuation, while the greater likelihood of being matched against a weaker opponent in the current stage increases it. Therefore, the nature of the game is affected by the rules for drawing opponents, such as seeding. A simulation of a two-stage game illustrates these issues.[5]

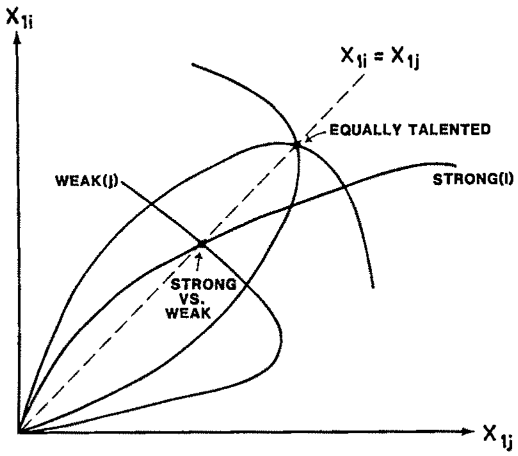


FIGURE 3

To simplify, assume two player types and constant elasticity cost and $h(x)$ functions. Type 1 is stronger than type 2 ($\gamma_1 > \gamma_2$). Since we have to keep track of each player's talent, the definition of β_s must be extended to

$$\beta_s(I,J) = P_s(I,J)\left[1-\mu P_s(J,I)\right], \tag{17}$$

where the P's are evaluated at equilibrium. Equation (17) includes (10) when $I = J$ because $P_s(I,I) = 1/2$.

Finals. Since $EV_1 - W_2 = \Delta W_1$ for all γ_I, symmetry of (3) implies $x_{1i} = x_{1j}$ irrespective of players' talents: $P_1(I,J) = \gamma_I/(\gamma_I + \gamma_J)$ in equilibrium. Effort is greater in a final match involving equally talented contestants than in one which matches a stronger against a weaker player (Figure 3). Using the same manipulations as before, we find

$$V_1(I,J) = \beta_1(I,J)\Delta W_1 + W_1. \tag{18}$$

Since $P_1(1,2) > P_1(2,1)$—a stronger player has a winning edge against a weak one in equilibrium, we have $\beta_1(1,2) > \beta > \beta_1(2,1)$.

Semifinals. Let π_1 denote the probability that the winner of the match in question

[4]The analysis of direct effort spillovers across stages is complicated by the fact that there may be asymmetric as well as symmetric pure-strategy equilibria. At symmetric equilibrium it is easy to show that fatigue and "burnout" requires more concentration on the top prize to penalize early-round "coasting." The force of "momentum" or learning requires less concentration at the top to maintain constant quality of play.

[5]Notice that a player is interested in what players in other matches are doing at any given stage because those outcomes determine who likely opponents will be in future stages. Equilibrium at each stage is a simultaneous 2^s player game: the problem does not disassemble pairwise and its complete solution must be simulated.

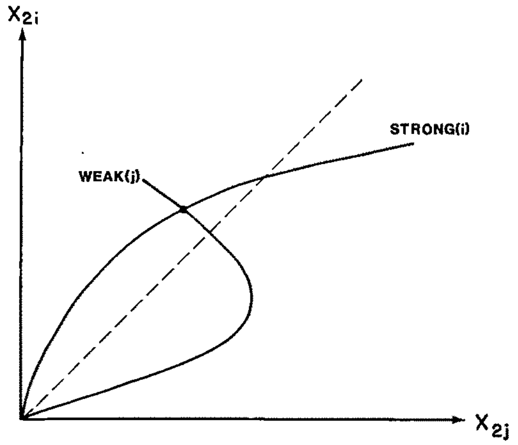


FIGURE 4

will confront a strong player in the finals. This depends on the identities and efforts chosen by players in the other match, but these actions are given to the opponents in this match in the Nash solution. Therefore,

$$(19)\quad EV_1(I) = [\pi_1\beta_1(I,1) + (1-\pi_1) \times \beta_1(I,2)]\Delta W_1 + W_2 = \tilde{\beta}_1(I)\Delta W_1 + W_2,$$

where

$$\tilde{\beta}_1(I) = \pi_1\beta_1(I,1) + (1-\pi_1)\beta_1(I,2)$$

and

$$(20)\quad EV_1(I) - W_3 = \tilde{\beta}_1(I)\Delta W_1 + \Delta W_2.$$

π_1 is smaller for the strong contestant implies $\tilde{\beta}_1(1) > \tilde{\beta}_1(2)$ because $\beta_1(1,2) > \beta_1(2,1)$.

There are two possible types of matches in the semi's. The equilibrium is symmetric if $I = J$, with $P_2(1,1) = P_2(2,2) = 1/2$. If $I \neq J$, the equilibrium is *not* symmetric because the stronger player has a greater value of continuation in (19) and (20). The strong player exerts greater efforts to win in equilibrium and $P_2(1,2) > \gamma_1/(\gamma_1+\gamma_2) = P_1(1,2)$ (see Figure 4). We find

$$V_2(I,J) = \beta_2(I,J)[\tilde{\beta}_1(I)\Delta W_1 + \Delta W_2] + W_3.$$

Furthermore,

$$(21)\quad \beta_2(1,2) > \beta_1(1,2) > \beta > \beta_1(2,1) > \beta_2(2,1),$$

which implies $V_2(1,2) > V_2(2,1)$ and $\tilde{\beta}_2(1) > \tilde{\beta}_2(2)$.

These formulas generalize for arbitrary s:

$$(22)\quad V_s(I,J) = \beta_s(I,J)[(\tilde{\beta}_1(I)\tilde{\beta}_2(I)\dots\tilde{\beta}_{s-1}(I))\Delta W_1 + (\tilde{\beta}_2(I)\dots\tilde{\beta}_{s-1}(I))\Delta W_2 + \dots + \Delta W_s] + W_{s+1}EV_{s-1} - W_{s+1}$$
$$= (\tilde{\beta}_1(I)\tilde{\beta}_2(I)\dots\tilde{\beta}_{s-1}(I))\Delta W_1 + (\tilde{\beta}_2(I)\dots\tilde{\beta}_{s-1}(I))\Delta W_2 + \tilde{\beta}_{s-1}(I)\Delta W_{s-1} + \Delta W_s$$

$$\tilde{\beta}_s(I) = \pi_s\beta_s(I,1) + (1-\pi_s)\beta_s(I,2),$$

where π_s is the probability a strong opponent will be encountered at s. An easy induction proves $\tilde{\beta}_s(1) > \tilde{\beta}_s(2)$ for all s, so (22) implies that the value of continuation is larger for stronger players at every stage of the game. The second expression in (22) also implies that a strong player works harder in a strong-weak match than a weak player does: the weak are eliminated with probability in excess of $\gamma_1/(\gamma_1+\gamma_2)$ at every stage except the last.

Since the value of the game is larger for stronger players, equilibrium in matches involving unequally talented players is asymmetric (except in the finals) and the definition of incentive maintenance must be extended. The most straightforward extension is a requirement that the same level of effort be maintained in all stages *within* any given type of pairing I against J: it is not feasible for effort to be maintained at a constant value *across* match types due to heterogeneity. Even this question cannot be answered in its entirety without additional structure, because the inequality in (21) cannot be extended in general. However, we have the following analytical result for the

TABLE 1—TWO-STAGE, TWO-TYPES SIMULATION ($\gamma_1 = 2, \gamma_2 = 1$)

	Spread: $\Delta W_1/\Delta W_2$				
	1	2	3	5	8
A. Semifinals ($s = 2$)					
$x_2(1,1)$	120.3	118.1	116.6	115.1	113.9
$x_2(2,2)$	92.6	76.4	60.7	55.5	47.7
$x_2(1,2)$	92.7	82.6	76.1	68.3	62.0
$x_2(2,1)$	81.6	66.7	57.7	47.4	39.6
$P_2(1,2)$	.69	.71	.73	.74	.76
B. Finals ($s = 1$)					
$x_1(1,1) = x_1(2,2)$	83.3	125.0	150.0	178.5	200.0
$x_1(1,2) = x_1(2,1)$	74.1	111.1	133.4	158.8	177.8
Pr(1,1)	.24/.48	.26/.51	.28/.53	.28/.55	.29/.57
Pr(2,2)	.04/.09	.04/.08	.03/.07	.03/.06	.03/.06
Pr(1,2)	.72/.43	.70/.41	.70/.40	.69/.39	.68/.37
C. Expected Total Effort					
Random	540.7	574.7	598.9	616.0	632.4
Seeds	507.4	537.2	554.4	573.3	586.9

Notes: Simulation for $\gamma_1 = 2$, $\gamma_2 = 1$, $\eta = \varepsilon = \mu = 1.0$. The term $x_s(I,J)$ is equilibrium effort expended by player of ability type I in match against opponent of ability type J when s stages remain; $P_2(1,2)$ is probability strong player wins semifinal round match against weak opponent; $P_1(1,2) = 2/3$ because final-round equilibrium is symmetric. Finals pairing probability $\Pr(I,J)$ is equilibrium probability of type I against type J in finals: First number refers to *random* initial draw, second number to strong/weak *seeds* in first round. The last rows give expected effort summed over all players in all matches and stages.

last two stages:

If the prize distribution is linear at the top ($\Delta W_1 = \Delta W_2$), effort by both players in strong-weak matches is larger in the semifinals than in the finals; and effort in matches between similar types is also larger in the semi's than in the finals.

The first part follows from the fact that $\tilde{\beta}_1(1)$ necessarily exceeds $\tilde{\beta}_1(2)$; while the second part follows from Section III (and in fact holds true for all stages when the prize structure is linear everywhere). The best reply for each player in any type of match is larger in the semi's than in the finals when $\Delta W_1 = \Delta W_2$. Consequently, the extra incremental prize at the top remains necessary to extend the horizon and help insure that the final match is the best match.

A small simulation for a two-stage game illustrates these ideas and shows some effects of seeding. To simplify the calculations, I chose $h(x) = c(x) = x$ (so $\varepsilon = \eta = \mu = 1.0$). Further, $\gamma_1 = 2$ and $\gamma_2 = 1$: true-to-form odds in a (1,2) match are 2-to-1 in favor of type 1. The simulation assumes that the game begins with two players of each type. The total purse is fixed at 1000 and $W_3 = 0$. The rank-prize structure is linear when $\Delta W_1/\Delta W_2 = 1$.

The first two rows of Table 1 show what might happen in a random draw which pulls strong-strong and weak-weak in the initial round. This happens half the time and guarantees a strong-weak final match pairing. Column 1 demonstrates that finals effort is smaller than semifinals effort when prizes are linear. Comparing across columns, we see that the final-round increment has to be quite large for strong players to exert more effort in the finals than in the semi's. The results are qualitatively similar for the other initial-round pairing possibility. These mixed matches would be assured by seeing, but occur only half the time with random draws. Notice that effort is smaller in mixed matches than in like matches, and that effort differences across player types are smaller in mixed matches. Strong players work very hard at round 1 to knock each other off and get into the finals when they know that their next opponent will be weak.

The probabilities of various final match-type pairings are shown in panel B. Neither seeding nor random draw guarantees that the best players survive to the finals, but seeding *doubles* the probabilities that they do. Under random draw, the most probable

(by far) final match is strong-weak. These probabilities are fairly insensitive to spread because the strong-player win-probability in a mixed first-round match is insensitive to the prize distribution with this parameter configuration. Notice that seeding makes a strong-strong final match the most probable outcome, but it comes at the cost of increasing the probability of a weak-weak final. However, this latter probability is small in either case.

Comparing across columns, we see that semifinals effort decreases and finals effort increases as the spreads grow larger. However, panel C shows that the second effect exceeds the first: expected total effort over all matches and stages increases with the spread. Most remarkably, total effort is greater when the initial draw is random than seeded. Seeding produces less variance in efforts in the first round, a lower mean in that round, and it most likely produces a better match among more talented opponents in the finals. The final interrank spread must be greatly elevated in the seeding game to produce expected total effort comparable to the no-seeding game. This suggests that seeds are observed when not simply total effort expended, but the distribution of the quality of play among players and stages, and guaranteeing the best match at the end, are important for the social productivity of the game. It justifies my reluctance to specify an additive social value function for the purposes of calculating an "optimal" prize structure.

V. Heterogeneous Contestants with Talents Unknown

Suppose we are interested in choosing the best out of T possible contestants. A round-robin design matches each player against every other and chooses the one with the largest overall win percentage. A sequential or knockout design eliminates a contender from further consideration after a certain number of losses. The sequential design promotes survival of the fittest and saves sampling costs by eliminating likely losers early in the game, but provides less precise information than the round-robin. The design choice comes down to comparing sampling costs with the value of more precision or the loss of making errors. H. A. David (1959, 1969) suggests that knockout designs have advantages over round-robins in selecting the best contestant, and Jean Gibbons, Ingram Olkin, and Milton Sobel (1977) prove it using sequential statistical decision theory. These issues are of great practical importance in medical trials. However, it is not possible to apply statistical decision theory alone to selection in human populations because no account is taken of contestants' incentives to optimize against the experimental design.

The main ideas are best illustrated in the case of "symmetric ignorance." Consider a sequential single elimination design, in which there are m types of contestants, all of whom share the same priors on the talents of others and who are equally ignorant about their own and others' talents. The distribution of types is common knowledge, and there is no private information. Estimates of own talents and the strength of the surviving field are updated as the game proceeds. This changing information feeds back into each contestant's strategy at every stage. When contestants have no more information about themselves than their surviving opponents do, it is clear that the interesting equilibrium is symmetric, because all survivors share the same information set—the same winning record, and choose the same strategy.

Let $\alpha_s(I)$ denote the probability that a player is type I when s stages remain to be played, and let $\tilde{\alpha}_s(J)$ denote the player's assessment that the current opponent is type J. Then, from Bayes' rule, the player's assessment of himself when $s-1$ stages remain, conditional on surviving (winning at stage s) is

$$\begin{aligned}(23)\quad \alpha_{s-1}(I) &= \Pr(\text{win at stage } s|I)\,\alpha_s(I)\Big/\Pr(\text{win at stage } s)\\ &= \alpha_s(I)\sum_J \tilde{\alpha}_s(J)P_s(I,J)\Big/\sum_I\sum_J \alpha_s(I)\tilde{\alpha}_s(J)P_s(I,J)\end{aligned}$$

where $P_s(I,J)$ is the win technology in (1); $\Sigma_J \tilde{\alpha}_s(J)P_s(I,J)$ is the conditional probability of winning given that one is type I. The denominator is the unconditional probability of winning at stage s. Since the initial prior is common, information is common at all stages, so $\alpha_s(I)=\tilde{\alpha}_s(I)$ in equilibrium. Furthermore, all contestants choose the same effort for given s, and $P_s(I,J)=\gamma_I/(\gamma_I+\gamma_J)$ in equilibrium: survival chances for each type run true to form at each stage. Finally, the unconditional equilibrium survival probability is always 1/2 in paired comparisons. In equilibrium (23) becomes

$$(24)\quad \alpha_{s-1}(I)=2\alpha_s(I)\times\sum_J \alpha_s(J)[\gamma_I/(\gamma_I+\gamma_J)].$$

Equation (24) implies survival of the fittest. To illustrate, suppose there are two types, with $\gamma_1>\gamma_2$. Let α_s be the expected proportion of stronger (type-1) players alive at s. Then (24) is

$$(25)\quad \alpha_{s-1}-\alpha_s=\alpha_s(1-\alpha_s)\omega,$$

where $\omega=(\gamma_1-\gamma_2)/(\gamma_1+\gamma_2)$ is the difference in form probabilities between types. The solution to (25) looks like a logistic. The weak are eliminated at the largest rate when $\alpha_s=1/2$, and are eliminated at a slower rate elsewhere. The rate of elimination of the weak also depends on ω. Convergence is very fast when ω is large. For example, if $\alpha_n=1/2$ and ω is close to unity (its maximum possible value) over 99 percent of expected survivors are strong after only three stages. More stages are required to select the fittest members of the population the smaller the initial values of α and ω.

In choosing a strategy a player must assess own and opponents' talents at each stage. The problem is illustrated for the case of two types, strong (γ_1) and weak (γ_2). We have

$$(26)\quad V_s(\alpha_s,\tilde{\alpha}_s)=\max\{\Pr(\text{win}|\alpha_s,\tilde{\alpha}_s)\times[V_{s-1}(\alpha_{s-1},\tilde{\alpha}_{s-1})-W_{s+1}]-c(x_{si})\},$$

where the win probability is conditioned on the information available at the beginning of stage s:

$$(27)\quad \Pr(\text{win}|\alpha_s,\tilde{\alpha}_s)=\alpha_s[\tilde{\alpha}_sP_s(1,1)+(1-\tilde{\alpha}_s)P_s(1,2)]+(1-\alpha_s)[\tilde{\alpha}_sP_s(2,1)+(1-\tilde{\alpha}_s)P_s(2,2)],$$

with $(\alpha_s,\tilde{\alpha}_s)$ updated according to (23). Thus in choosing x_s the player weighs the possibilities of own and opponent's talent pairings by the information currently available. This information depends on past data, exogenous as of stage s. The player's assessment of the future strength of an opponent, $\tilde{\alpha}_{s-1}$, depends on the given efforts of players in other matches. However, the player's assessment of his own talent in the next stage depends on today's actions and outcomes, from (23), and this (the value of information) also enters the calculation for choice of x_s. The Bayesian link between stages s and $s-1$ introduces an interstage linkage in strategies that is not present when talents are known.

The first-order condition for this problem is

$$(28)\quad \frac{\partial\Pr(\text{win}|.)}{\partial x_{si}}[V_{s-1}(\alpha_{s-1},\tilde{\alpha}_{s-1})-W_{s+1}]+\Pr(\text{win}|.)[\partial V_{s-1}(\alpha_{s-1},\tilde{\alpha}_{s-1})/\partial\alpha_{s-1}]\times(\partial\alpha_{s-1}/\partial x_{si})-c'(x_{si})=0.$$

The derivative $\partial\Pr(\text{win}|.)/\partial x_{si}$ is calculated from (27) and $\partial\alpha_{s-1}/\partial x_{si}$ is calculated from (23), both given x_{sj}. An expression for the information term $\partial V_{s-1}/\partial\alpha_{s-1}$ is found by applying the envelope property to (26):

$$(29)\quad \partial V_s(\alpha_s,\tilde{\alpha}_s)/\partial\alpha_s=[V_{s-1}(\alpha_{s-1},\tilde{\alpha}_{s-1})-W_{s+1}]\times(\partial\Pr(\text{win}|\alpha_s,\tilde{\alpha}_s)/\partial\alpha_s).$$

The symmetric solution is characterized by (28) evaluated at $\alpha_s = \tilde{\alpha}_s$ and $x_{si} = x_{sj}$ for all s.

Writing $V_s = V_s(\alpha_s, \alpha_s)$ detailed calculations at equilibrium yield

$$\text{(30a)} \quad \partial V_s / \partial \alpha_s = (V_{s-1} - W_{s+1})(\omega/2);$$

$$\text{(30b)} \quad \partial \Pr(\text{win}|.)/\partial x_{si} = (h'/h) \times [(1/4) - \alpha_s(1-\alpha_s)(\omega^2/2)];$$

$$\text{(30c)} \quad \partial \alpha_{s-1}/\partial x_{si} = -\alpha_s(1-\alpha_s) \times \omega\left[\left(\omega\alpha_s - \frac{\gamma_1}{\gamma_1+\gamma_2}\right)^2 + \frac{\gamma_1\gamma_2}{(\gamma_1+\gamma_2)^2}\right].$$

Condition (30a) shows that the value of continuation is increasing in own-assessment of talent, and that its incremental value is increasing in ω, the difference in form probabilities. The value of information is small when contestants are not very different from each other. The marginal effect of effort on winning (in (30b)) is decreasing in population heterogeneity (ω) and in the uncertainty with which players assess themselves at each stage (α). Uncertainty is a force that dampens incentives to perform and is greatest at $\alpha_s = 1/2$. This effect disappears as uncertainty is resolved. Equation (30c) shows that greater effort *reduces* the posterior assessment of strength.[6] Given the equilibrium effort of the opponent, the winning contestant is more probably of greater talent if less effort has been expended. The elimination design places extra value on strength, and private incentives to experiment to discover own strength is another force tending to make players hold back efforts at earlier stages. However, this term also vanishes as uncertainty is resolved.

Substituting (30) in (28) and manipulating into elasticity form, we have

$$\text{(31)} \quad c(x_s) = [(\mu/4) - A_s](V_{s-1} - W_{s+1}) - B_s(V_{s-2} - W_s),$$

where

$$\text{(32)} \quad A_s = \mu\alpha_s(1-\alpha_s)(\omega^2/2)$$

$$B_s = \mu(\omega^2/4)\alpha_s(1-\alpha_s) \times \left[\left(\omega\alpha_s - \frac{\gamma_1}{\gamma_1+\gamma_2}\right)^2 + \frac{\gamma_1\gamma_2}{(\gamma_1+\gamma_2)^2}\right] \quad \text{for } s \geq 2$$

$$B_1 = 0$$

The boundary condition $B_1 = 0$ holds because information has no value in the finals. Equation (25) is used to calculate A_s and B_s. Substituting into the value function and subtracting W_{s+2} provides a recursion for the increments $V_s - W_{s+2}$ in (31):

$$\text{(33)} \quad V_s - W_{s+2} = (\beta - A_s)(V_{s-1} - W_{s+1}) + B_s(V_{s-2} - W_s) + \Delta W_{s+1},$$

with boundary condition $V_0 - W_2 = \Delta W_1$.

Conditions (31) and (33) plus the calculation of A_s and B_s from (32) and (25) represent the complete solution of the symmetric ignorance problem. Notice that this solution converges in the limit to that of equal known talents (Section III) as α_s approaches unity, because A_s and B_s go to zero. Hence the extra increment in the final interrank spread is required for incentive maintenance in a sufficiently long game, irrespective of the initial distribution of talents.[7] By a similar

[6] Updating own-assessment of talent conditional on losing has no value because losers are eliminated. It has value in games with double or more eliminations, but equilibrium is not symmetric. Nor is it symmetric if contestants observe finer information than each player's previous win-loss record.

[7] The remaining case is one of private information, where each player knows his own type, but has prob-

token, the earlier result holds approximately when heterogeneity is small.

In fact, heterogeneity must be quite large for the value of information to have much effect on the incentive maintaining prize structure of Figure 2. For example, consider the case where $\gamma_1 = 2$, $\gamma_2 = 1$ and $\mu = 1$. The strong type wins two-thirds of the time and $\omega = 1/3$. Direct calculation reveals that B_s is of order 10^{-3}, and A_s is of order 10^{-2}. Therefore the second difference effects in (31) and (33) are negligible and Figure 2 is a very close approximation to the incentive maintaining prize structure. When the strong player wins three-fourths of the time, the corresponding orders of magnitude are 10^{-2} for both terms, so the approximation in Figure 2 remains very good: there are only a few minor wiggles.

Significant departures from Figure 2 occur when there are major differences between types, but this is mainly due to the incentive dilution effects of uncertainty. Even when the strong player wins 90 percent of the time, the terms in B_s remain of order 10^{-2} and the second difference (value of information) terms are negligible. But the terms in A_s show more variation with α_s. The term $(\beta - A_s)$ is smallest in those stages where uncertainty is largest. The interrank spread must be increased in those stages for x_s to be maintained, to overcome larger discounting of the future. Thus in a tournament where the proportion of strong players is relatively small in the first round, early-round incentive-maintaining prizes are approximately linear because there is little uncertainty. As the weak players are eliminated and α_s rises toward $1/2$, uncertainty is *increasing* and the interrank spread has to increase to overcome this effect. If the game is long enough for α_s to exceed $1/2$, uncertainty is decreasing and interrank spreads are decreasing for incentive maintenance. They increase again toward the end, due to the horizon effects. If the initial field is equally split ($\alpha_N = 1/2$), resolution-of-uncertainty acts to distribute the prize money more equally across the ranks. If the initial proportion α_N is small and the game is long, incentive-maintaining prizes redistribute from the extremes toward the middle.

Finally, the expected selection recursions in (24) show that the social value of information is independent of x_s in the symmetric equilibrium: all information in selecting strong players for survival is embedded in the elimination design itself, and incentives for contestants to produce private information come to naught. The attempt by all players to gain informational advantages in calculating their private strategies cancel each other out because of the ordered quality of competition. No one obtains an informational edge over that inherent in the design. There is a role for the prize structure to discourage these socially useless actions, and this requires less concentration of the prize money at the top, to reduce the private value of information. However the calculations above suggest that these effects are relatively minor unless differences in talents are large.

VI. Conclusions

The chief result is identifying a unique role for top-ranking prizes to maintain performance incentives in career and other games of survival. Extra weight on top-ranking prizes is required to induce competitors to aspire to higher goals independent of past achievements. There are many rungs in the ladder to aspire to in the early stages of the game, and this plays an important role in maintaining one's enthusiasm for continuing. But after one has climbed a fair distance, there are fewer rungs left to attain. If top prizes are not large enough, those that have succeeded in achieving higher ranks rest on their laurels and slack off in their attempts to climb higher. Elevating the top prizes effectively lengthens the ladder for higher-ranking contestants, and in the limit makes it appear of unbounded length: no matter how far one has climbed, there is

abilistic assessments of opponents' types. Then Bayesian updating applies to opponents only. The analysis of this case is conceptually straightforward, but the equilibria are not symmetric and few analytical results are available. It is omitted for that reason. Still, the result on concentration of the purse on the top applies because survivors at the last few stages are relatively homogeneous.

always the same length to go. In examining the relation between wages and marginal products, the concept of marginal productivity must be extended to take account of the value to the organization of maintaining incentives and selecting the best personnel to the various rungs, not only the contribution at each step. Payments at the top have indirect effects of increasing productivity of competitors further down the ladder.

There is another interesting class of questions in this type of competition. Adam Smith held the opinion that there is natural tendency for competitors to overestimate their survival chances ("overweaning conceit"), while Alfred Marshall held the opposite opinion. Further analysis shows how biased assessments of talent affect survival. There is a clear disadvantage to pessimism and underestimation of own talents. The pessimist doesn't try hard enough because opponents appear relatively stronger, and also because the true value of continuation is underrated. An elimination design is disadvantageous to the timid. They do not survive very long. The effects of overestimation and optimism are more complicated. For strong players and among any contestants in a field of comparable types, optimism has two effects: the optimist has a tendency to slack off due to underestimation of the relative strengths of the competition, but overestimates the own-value of continuation, which induces greater effort. Optimism has no clear-cut effects on altering survival probabilities. However, the second effect vanishes in the finals, and winning chances are reduced. Optimism has positive survival for weak players in a strong field. A weaker player who feels closer to the average field strength than is true, works harder on both counts and is not eliminated as quickly as another weak competitor with more accurate self-assessments.

When contestants' abilities are unknown, private incentives to optimize against the design for personal informational advantage lead to socially useless actions. These in the end do not produce any more information than is already embodied in the game itself and must be discouraged by concentrating less of the purse at the top. There are also private incentives for a contestant to invest in signals aimed at misleading opponents' assessments. It is in the interest of a strong player to make rivals think his strength is greater than it truly is, to induce a rival to put forth less effort. The same is true of a weak player in a weak field. However, it is in the interests of a weak player in a strong field to give out signals that he is even weaker than true, to induce a strong rival to slack off. Weighting the top prizes less heavily reduces these inefficient signaling incentives.

REFERENCES

Becker, Gary S. and Stigler, George J., "Law Enforcement, Malfeasance and the Compensation of Enforcers," *Journal of Legal Studies*, January 1984, *3*, 27–56.

Beckmann, Martin J. *Rank in Organizations*, Berlin: Springer-Verlag, 1978.

Bronars, Stephen, "Underdogs and Frontrunners: Strategic Behavior in Tournaments," Texas A&M University, 1985.

Carmichael, H. Lorne, "The Agent-Agents Problem: Payment by Relative Output," *Journal of Labor Economics*, January 1983, *1*, 50–65.

David, H. A., "Tournaments and Paired Comparisons," *Biometrika*, June 1959, *46*, 139–49.

______, *The Method of Paired Comparisons*, London: Charles Griffen and Co., 1969.

Gibbons, Jean D., Olkin, Ingram and Sobel, Milton, *Selecting and Ordering Populations: A New Statistical Methodology*, New York: Wiley & Sons, 1977.

Green, Jerry R. and Stokey, Nancy L., "A Comparison of Tournaments and Contracts," *Journal of Political Economy*, June 1983, *91*, 349–65.

Holmstrom, Bengt, "Moral Hazard in Teams," *Bell Journal of Economics*, Autumn 1982, *13*, 324–40.

Lazear, Edward P., "Why Is There Mandatory Retirement?," *Journal of Political Economy*, October 1981, *89*, 841–64.

______ **and Rosen, Sherwin,** "Rank Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, October 1981, *89*, 841–64.

Loury, Glenn C., "Market Structure and In-

novation," *Quarterly Journal of Economics*, August 1979, *94*, 395–410.

Malcomson, James M., "Work Incentives, Hierarchy, and Internal Labor Markets," *Journal of Political Economy*, June 1984, *92*, 486–507.

Murphy, Kevin J., "Ability, Performance and Compensation: A Theoretical and Empirical Investigation of Labor Market Contracts," unpublished doctoral dissertation, University of Chicago, 1984.

Nalebuff, Barry J. and Stiglitz, Joseph E., "Prizes and Incentives: Toward a General Theory of Compensation and Competition," *Bell Journal of Economics*, Spring 1984, *2*, 21–43.

O'Keeffe, Mary, Viscusi, W. Kip and Zeckhauser, Richard J., "Economic Contests: Comparative Reward Schemes," *Journal of Labor Economics*, January 1984, *2*, 27–56.

Rosen, Sherwin, "Authority, Control and the Distribution of Earnings," *Bell Journal of Economics*, October 1982, *13*, 311–23.

Rosenbaum, James E., *Career Mobility in a Corporate Hierarchy*, Orlando: Academic Press, 1984.

The Architecture of Economic Systems: Hierarchies and Polyarchies

By RAAJ KUMAR SAH AND JOSEPH E. STIGLITZ*

This paper presents some new ways of looking at economic systems and organizations. Individuals' judgments entail errors; they sometimes reject good projects and accept bad projects (or ideas). The architecture of an economic system (i.e., how the decision-making units are organized together within a system, who gathers what information, and who communicates what with whom) affects the errors made by individuals within the system, as well as how those errors are aggregated.

There is a widespread belief that the performance of an economic system or organization is influenced by its internal structure. In this paper, we present some new ways of looking at the relationship between performance of an economic system and certain aspects of its structure, which we refer to as its *architecture*. The architecture (like that of a computer or electrical system) describes how the constituent decision-making units are arranged together in a system, how the decision-making authority and ability is distributed within a system, who gathers what information, and who communicates what with whom.

The two specific architectures studied in this paper are called polyarchies and hierarchies. We think of a *polyarchy* as a system in which there are several (and possibly competing) decision makers who can undertake projects (or ideas) independently of one another. In contrast, decision-making authority is more concentrated in a *hierarchy* in the sense that only a few individuals (or only one individual) can undertake projects while others provide support in decision making. These two architectures are suggestive of a market-oriented economy and a bureaucracy-oriented economy, respectively.

The aspect of organizational performance on which we focus is the quality of decision making. All individuals make errors of judgment: some projects that get accepted should have been rejected, and some projects are rejected that should have been accepted. Using an analogy from the classical theory of statistical inference, these errors correspond to Type-II and Type-I errors.

How individuals are arranged together affects the nature of the errors made by the economic system. For example, in a market economy, if one firm rejects a profitable idea (say, for a new product), there is a possibility that some other firm might accept it. In contrast, if a single bureau makes such decisions and this bureau rejects the idea, then the idea must remain unused. The same, however, is also true for those ideas that are unprofitable. As a result, one would expect a greater incidence of Type-II errors in a polyarchy, and a greater incidence of Type-I errors in a hierarchy.

The costs of acquiring and communicating information (leading to misjudgments by individuals) are the central features of the technology underlying our analysis. These costs include the direct costs (time and re-

*Yale University, New Haven, CT 06520, and Princeton University, Princeton, NJ 08544, respectively. Financial support from the National Science Foundation and the Hoover Institution, Stanford is gratefully acknowledged. An earlier version of this paper was presented at seminars at Berkeley, Chicago, Columbia, Minnesota, Pennsylvania, Rutgers, San Diego, Stanford, and Yale, and to the 1984 European meetings of the Econometric Society in Madrid. We are indebted to the participants in those seminars for their insightful comments. We are especially grateful to John Geanakoplos, Alvin Klevorick, Mark Machina, James March, Paul Milgrom, Barry Nalebuff, Michael Rothschild, and an anonymous referee for helpful suggestions.

sources) and the indirect costs that result from the inevitable contamination that occurs in the process of information communication. Communication, like decision making, is always imperfect. No individual ever fully communicates perfectly what he knows to another.

Another important feature is the limited capabilities of individuals to gather, absorb, and process information within a limited amount of time. This is why organizations, groups of individuals, may be able to do more (make better decisions) than any single individual. But the fact that communication is costly and imperfect means that an organization with two individuals, each of whom can process a given amount of information in, say, a month, is not the same as a single individual who has the capacity of processing twice that amount of information within the same time period.

The paper is organized as follows. In Section I, we present a simple model of the decision structure within a polyarchy and a hierarchy. In Section II, we assume that the nature of an individual's errors and the mix of available projects is exogenously specified, and analyze how changes in these exogenous features influence the relative performance of the two systems. In Section III, we discuss the collection and processing of information in the two systems. In particular, we show how (Bayesian) screening rules are determined, and how the two system's performances compare with these endogenously determined individuals' errors. Section IV discusses briefly some extensions of the analysis, while Section V discusses alternative interpretations and applications. Proofs of most results are omitted for brevity; these are available in our 1985a working paper.

I. The Basic Model

The problem facing the economic systems under study is to choose which of a set of projects to undertake. Each project has a net benefit, x,[1] that can be positive or negative. There are N available projects. The density function of projects is given by $g(x)$.

The task of individuals within the organization is to evaluate ("screen") the projects. We assume that the only feasible communication is whether the project is, in the judgment of the evaluator, "good" or "bad," that is, whether it should be accepted (or passed on for further evaluation) or rejected. (For now, we can think of a screener as a black box that flashes a light when it deems a project to be good.) The probability that a given individual judges a project to be good, p, is a function of its quality. We call the function $p(x)$ the *screening function*. It can take any form, provided $1 \geq p(x) \geq 0$, for all x, and the strict inequalities hold for at least some x.

Two properties of the screening function are of special interest. The first is its slope, $p_x(x)$.[2] We assume $p_x(x)$ is positive, that is, a project with higher profit has a higher local probability of being accepted by a screen. Further, if p and p^1 represent two screens, and if $p_x(z) > p^1_x(z)$, then we refer to the former screen as locally more *discriminating* at $x = z$. The second important property of screens is the level of $p(x)$. If $p(z) > p^1(z)$, then we call the former screen locally *slacker*, and the latter locally *tighter*, at $x = z$. An example is the linear screening function, for which $p(x)$ can be expressed as

$$(1) \qquad p(x) = p(\mu) + p_x(x - \mu),$$

where $\mu = E[x]$ is the mean of the initial portfolio. Clearly, a higher $p(\mu)$ and p_x imply globally higher slackness and discriminating capability.

If screening were perfect, then the architecture of a system has no effect on its output because all projects with $x > 0$ would

[1] This scalar valuation includes all relevant benefits and costs. Also, we are assuming that the interproject externalities are not significant (i.e., the profit from one project does not depend significantly on whether some other projects are undertaken or not), and that there is no restriction on the number of projects that can be undertaken.

[2] A letter subscript denotes the variable with respect to which a partial derivative is being taken.

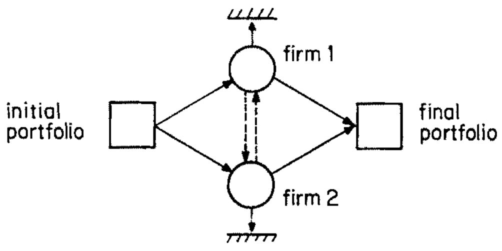


FIGURE 1. POLYARCHY

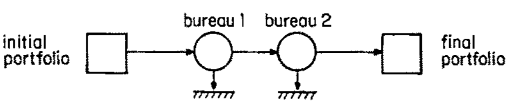


FIGURE 2. HIERARCHY

be accepted and those with $x<0$ would be rejected; that is, $p(x)=1$ if $x>0$, and $p(x)=0$ if $x\leq 0$. Without perfect screening, the architecture of the economic system determines the conditions under which a project gets selected and, hence, it affects the system's output. In the following simple model, we consider a polyarchy consisting of two firms, and a hierarchy consisting of two bureaus.

The decision process in a polyarchy and a hierarchy are depicted in Figures 1 and 2, respectively. In a polyarchy, the two firms screen the projects independently. For specificity, one may think of projects arriving randomly (with probability one-half) at one of the two firms. If a particular project is accepted by a firm, then it is no longer available to the other firm. If the project is rejected, then it goes to the other firm where, once again, it can be accepted or rejected (but firms cannot tell which of the projects that they are evaluating have been previously reviewed). Neither firm screens the same project twice, so that a project cannot cycle back and forth between firms. The portfolio of projects selected in a polyarchy therefore consists of the projects accepted separately by each of the two firms.

In contrast, in a hierarchy, all projects are first evaluated by the lower bureau (bureau 1); those that are accepted are forwarded to the higher bureau (bureau 2) and others are discarded. The projects selected by the system then are those which are selected by the higher bureau. Drawing an analogy from the design of relay circuits, the screens are placed in *series* in a hierarchy, whereas they are placed in *parallel* in a polyarchy.

In a polyarchy, the probability that a project that goes to the first firm is approved is $p(x)$. It gets rejected with probability $1-p(x)$; the probability that it then gets approved by the second firm is again $p(x)$. Hence the total probability of acceptance is $p(x)+(1-p(x))p(x)=p(x)(2-p(x))$. Similarly, in a hierarchy, the probability that a project is approved by the lower bureau is $p(x)$. The probability that the same project given to the higher bureau is approved is again $p(x)$. Hence, the probability of a project being approved is $p^2(x)$.

The probability that the project x will be accepted in the system s is denoted by $f^s(x)$, where the superscripts $s=P$ and H represent a polyarchy and a hierarchy, respectively. If individuals' decisions are independent across screens and projects, then

$$(2) \qquad f^H\equiv\left(p^H\right)^2; \qquad f^P\equiv p^P\left(2-p^P\right).$$

In the comparison of the two systems below, we assume that both systems face the same set of available projects, and that they have the same screening function. The latter assumption is dropped in Section III.

II. Comparative Performance with Identical Screening Functions

We investigate two questions in this section: what is the relative performance of the two systems, and how is it affected by the properties of the available project portfolio, and the screening function?

A. *The Size of Final Portfolios*

The proportion of the initial portfolio selected by the two systems, n^s, is just $\int f^s(x)g(x)dx\equiv E[f^s]$. Denoting the differ-

ence in these proportions by Δn, we find that[3]

$$\Delta n = n^P - n^H > 0, \tag{3}$$

since, from (2), $f^P - f^H = 2p(x)\{1-p(x)\} \geq 0$ for all x, and it is strictly positive for some x.

PROPOSITION 1: *A polyarchy selects a larger proportion of the available projects than does a hierarchy.*

The reason behind this result is intuitive. Consider a hypothetical situation in which the second firm in a polyarchy does not exist, and the higher bureau in a hierarchy does not exist. The proportion of projects accepted in the two systems would then be the same, namely, $E[p(x)]$. Since the second firm accepts at least some projects, and since the higher bureau rejects at least some projects, the actual proportion of projects accepted in a polyarchy must exceed that in a hierarchy. It is also obvious that this result holds for good as well as bad projects. Further, the result does not depend on how one defines good vs. bad projects, provided there is some probability that a screen will accept at least some good and some bad projects. It immediately follows that: *A polyarchy accepts a larger proportion of good as well as bad projects compared to a hierarchy, no matter how one defines good and bad projects.* Therefore, *the incidence of Type-I error is relatively higher in a hierarchy, whereas the incidence of Type-II error is relatively higher in a polyarchy.*

The above result suggests that there may be circumstances in which a polyarchy performs better than a hierarchy (when it is more important to avoid Type-I errors) and other circumstances in which a hierarchy performs better than a polyarchy (when it is more important to avoid Type-II errors).

To determine the impact of initial portfolios on the size of final portfolios, note from (2) that $f^s(x)$ is increasing in x. Additionally, $f^P(x)$ is concave and $f^H(x)$ is convex in x, if the screening function is linear. Therefore, the standard properties of statistical dominance (under an assumption that the end points of the projects' distribution are fixed) yield the following result.

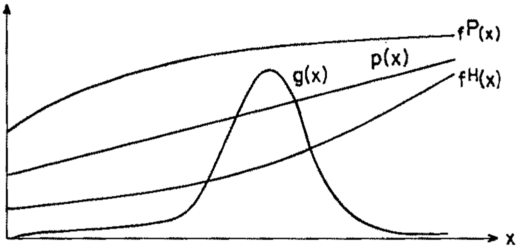


FIGURE 3. PROBABILITIES OF ACCEPTANCE IN ALTERNATIVE SYSTEMS

PROPOSITION 2: *A worsening in the initial portfolio in the sense of first-order stochastic dominance leads to a smaller proportion of initial projects being selected in both systems. With a linear screening function, a mean-preserving spread in the initial portfolio leads to a smaller proportion of initial projects being selected in a polyarchy, and a larger proportion being selected in a hierarchy.*

These results can be seen in Figure 3. As shown, f^P and f^H are concave and convex in x, since $p(x)$ in linear. n^s is the area above the x-axis bounded by the product of f^s and g. Naturally, this area corresponding to f^P is larger than that corresponding to f^H; and this area enlarges, for both a polyarchy and a hierarchy, if the density weight shifts from lower x to higher x. Also, if the density weight shifts from the mean to the two sides, due to a mean-preserving spread, then the area representing n^s decreases in a polyarchy and it increases in a hierarchy.

Straightforward calculations allow one to ascertain how n^s is influenced by changes in the two parameters of the linear screening function. We find the following.

PROPOSITION 3: *With a linear screening function, a higher slackness in screening*

[3]Here as elsewhere, N, the number of available projects plays no role, and we therefore suppress it. The variables representing the performance of a system are thus normalized by the number of available projects.

$(p(\mu))$ raises the proportion of projects selected in both systems. And, a higher discriminating ability in screening (p_x) lowers the proportion selected in a polyarchy, whereas it raises the proportion selected in a hierarchy.

B. *Profits in Alternative Systems*

Two Types of Projects. In the case where the set of projects to be reviewed consists only of two types of projects, we can obtain a complete characterization of the conditions under which the (expected) output is higher under polyarchy or hierarchy. The initial portfolio is represented by the return on good projects, $z_1 > 0$; the return on bad projects, $-z_2 < 0$; and the proportion of good projects, α. The screening function is characterized by the probability that a good project gets accepted, denoted by $p_1 = p(3_1)$, and the probability that a bad project is accepted, denoted by $p_2 = p(-z_2)$. If $Y^s = E[xf^s]$ denotes the output, and $\Delta Y = Y^P - Y^H$ denotes the difference between the outputs of the two systems, then

$$\Delta Y = 2z_2(1-\alpha) \times [ap_1(1-p_1) - p_2(1-p_2)], \tag{4}$$

where $a = z_1\alpha/z_2(1-\alpha)$ is a summary representation of the quality of the initial portfolio.

An improvement in the initial portfolio in the present model is represented by a larger a (i.e., a larger α or a larger z_1/z_2). It follows from (4) that a worse initial portfolio implies that the relative performance of a polyarchy, compared to a hierarchy, is worse. This is simply because the relative advantage of a hierarchy is in rejecting bad projects, whereas the relative advantage of a polyarchy is in accepting good projects. If the initial portfolio worsens, then the former advantage becomes increasingly more important and the relative performance of a hierarchy improves. However, we must caution that the probability that a project is accepted or rejected by a screen (i.e., the rules for project acceptance and rejection), might be affected by the mix of available projects, among other things. One might suspect, for instance, that if there was a large proportion of bad projects, screening would become relatively tighter in a polyarchy, and this might improve its relative performance. Endogenous screening functions with such properties are discussed later.

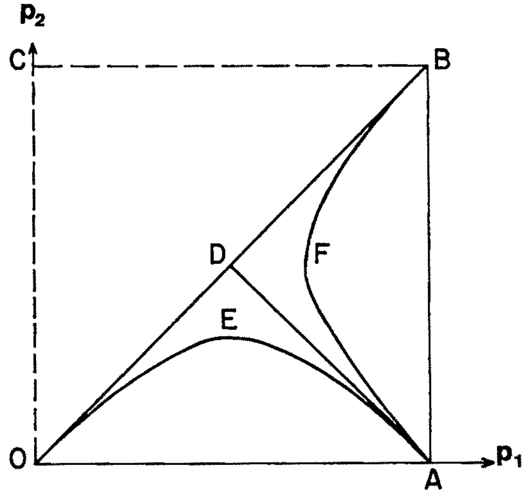


FIGURE 4. COMPARISON OF A POLYARCHY AND A HIERARCHY

The above expression also allows us to demarcate the parameter space into two regions: one in which a polyarchy has a higher output than a hierarchy, and the other in which the reverse holds.

Figure 4 summarizes the results. We are concerned only with the area below the 45° line, since screens have some discriminating capability; that is, $p_1 > p_2$.

First consider the case where the initial portfolio is moderately good; that is, $a = 1$. This happens, for instance, if the initial portfolio has equal number of good and bad projects ($\alpha = 1/2$) and if the gains and losses from the two types of projects are symmetric ($z_1 = z_2$). In this case, a polyarchy has a higher output if

$$1 - p_1 > p_2. \tag{5}$$

In Figure 4, thus, a polyarchy performs better in the area *ODA* and the reverse holds in the area *ADB*.

This result has a simple explanation. Recall that $(1-p_1)$ is a screen's Type-I error, the probability of rejecting a good project; and p_2 is a screen's Type-II error, the probability of accepting a bad project. Now, if a screen is more likely to reject a good project than to accept a bad project, that is, if (5) holds, then it must be the case that a polyarchy (which gives a second chance to the rejected projects) would do better.

If the initial portfolio is worse, that is, $a<1$, then, from (4), we find that the parameter space is separated by a hyperbola like *OEA*, which is inside the region *ODA*. A polyarchy has a higher profit within the region *OEA*, and the reverse holds outside of it. The region *OEA* shrinks as the initial portfolio becomes worse, and it coincides with the line *OA* if $a \approx 0$. The opposite case, in which the initial portfolio is better (i.e., $a>1$) has a parallel implication. A polyarchy then has a higher profit outside of the region *AFB*, and the reverse holds inside it.

There is another way in which the results can be seen intuitively. Suppose that we subjected each project to two screenings. Clearly, if both screens indicated that the project was bad, the project should be rejected, and if both indicated that the project was good, it should be accepted. A tradeoff arises in those cases where there is a mixed review. Whether a project with a mixed review should be undertaken depends on the profit from such a project. The probability of a good project getting a mixed review is $2p_1(1-p_1)$, while the probability of a bad project getting a mixed review is $2p_2(1-p_2)$. Hence the expected profit from projects with mixed reviews is the same as (4). Now, if it turns out that the expression (4) is positive, it means that projects with mixed reviews should be accepted; this is precisely what polyarchy ensures. Similarly, if it turns out that (4) is negative, it means that the projects with mixed reviews should be rejected, and this is precisely what hierarchy ensures.

A General Project Portfolio. Before concluding this subsection, we briefly consider an initial portfolio consisting of a continuum of projects. Recall from (2) that $f^P - f^H = 2p(1-p)$. Then, $\Delta Y = 2E[x\psi]$, where $\psi = p(1-p)$. To determine the effect of a change in the initial portfolio, we note

$$(6) \qquad \psi_x = (1-2p)p_x;$$
$$\psi_{xx} = -2p_x^2 + (1-2p)p_{xx}.$$

Since ψ can be either a concave or convex function of x, the effect of a mean-preserving change in the initial portfolio is, in general, ambiguous. If the range of x is small, however, then using (6), we obtain

$$(7) \quad E[x\psi] \approx \psi(\mu)\mu + \psi_x(\mu)E[x(x-\mu)]$$
$$= p(\mu)((1-p(\mu))\mu + (1-2p(\mu))$$
$$\times p_x(\mu)E\left[(x-\mu)^2\right].$$

Hence, if $\mu \geq 0$, and $p(\mu) < \frac{1}{2}$, then a polyarchy has a larger output than hierarchy; further, an increase in the variance of the portfolio improves the relative performance of polyarchy, regardless of the value of μ.

When the screening function is linear, then additional comparative statics results can be easily obtained. For instance, if the initial portfolio contains projects symmetrically distributed around zero mean, then a polyarchy performs better or worse than a hierarchy depending simply on whether $p(\mu)$ is less than or more than one-half, that is, whether the screening is tight or slack. A higher mean or a greater negative skewness of the initial portfolio, on the other hand, improves the relative performance of a polyarchy.

III. Endogenous Screening Rules

The individual decision makers in the above model can be interpreted to be Bayesian, with each of them receiving a binary (imperfect) signal concerning the quality of the projects. More generally, individuals observe a much richer set of signals that they have to interpret; they have to decide, in other words, under what conditions they will recommend that the project be undertaken. Assume, for instance, that the project eval-

uator observes

$$y = x + \theta. \tag{8}$$

Project evaluators use reservation levels for screening: a project is accepted if its observed profit is above the reservation level, R, and it is rejected otherwise.[4] Assume θ is distributed independently of x and denote the distribution function of θ by $M(\theta)$ and its density by $m(\theta)$. The screening function, then, is given by

$$p(x,R) \equiv Prob[y \geq R] = 1 - M(R - x). \tag{9}$$

The above expression yields $p_x \geq 0$, and $p_R \leq 0$: the probability that a project is accepted by a screen is increasing in the quality of the project, and it is decreasing in the reservation level. Increasing R increases the probability of a good project being rejected (Type-I error) and decreases the probability of a bad project being accepted (Type-II error). The reservation level R is chosen to balance off these errors.

In a polyarchy, denote the two firms by superscripts i and j. For firm i, R^i is the reservation level, $p^i \equiv p(x, R^i)$ is the screening function, and Y^{iP} denotes the output; then

$$Y^{iP} = E[xp^i(2-p^j)]/2; \quad Y^P = Y^{1P} + Y^{2P}. \tag{10}$$

We now turn to a comparison of the reservation levels under the two systems, and using these results, to a comparison of performance. The reservation level in a hierarchy, R^H, maximizes $Y^H = E[xp^2]$; that is, it satisfies

$$Y_R^H = 2E[xpp_R] = 0. \tag{11}$$

To emphasize the independence (and potential competition) between the two firms in a polyarchy, we assume that their reservation levels are determined without coordination. We focus on the symmetric Nash optimum; from (10), the corresponding reservation level, R^P, is characterized by

$$E[x(2-p)p_R] = 0. \tag{12}$$

As a benchmark, we also note that in the case of a *coordinated polyarchy* (where the reservation level for the firms is set to maximize the combined output Y^P), the reservation level, R^C, is obtained by equating

$$Y_R^P = 2E[x(1-p)p_R] \tag{13}$$

to zero. Using expressions (11) to (13), one can show the following:[5]

PROPOSITION 4: $R^C > R^P > R^H$. *That is, the screening in a polyarchy is more conservative than that in a hierarchy, but less conservative than that in a coordinated polyarchy.*

This result has an intuitive explanation. While in a hierarchy, the lower bureau knows that its decisions are rechecked at the upper bureau; and the upper bureau knows that all projects it receives have been checked at the lower bureau; in a polyarchy, each firm knows that its decision will not be rechecked; and to make matters worse, it knows that the set of projects which it is examining includes many that have already been examined elsewhere, and have been rejected. This conservatism is reflected in market economies by firms insisting on a high "expected" re-

[4] The optimal policies can be characterized in terms of a single reservation level only if some mild regularity conditions are satisfied by the nature of the error terms.

[5] To show this, we define $c(x) = (1-p(x))/p(x)$, and observe that the turning points as well as global maximum of $c(0)Y^H$ are identical to those of Y^H for any constant $c(0)$, and that $d(c(0)Y^H - Y^P)/dR = 2Expp_R[c(0) - c(x)] < 0$, since $c_x(x) < 0$ and $p_R < 0$. Assume to the contrary: that $R^H > R^C$. Then from $Y_R^P > c(0)Y_R^H$, it follows that $Y^P(R^H) - Y^P(R^C) > c(0)\{Y^H(R^H) - Y^H(R^C)\}$. The left-hand side is negative and the right-hand side is positive. This contradiction establishes that $R^C > R^H$. The uncoordinated polyarchy can be thought of as maximizing $Ex[2p - .5p^2]$. The result that $R^C > R^P > R^H$ follows along exactly parallel lines to our proof that $R^C > R^H$.

turn in order to undertake a project. For example, firms often have a decision rule that only projects with an expected return in excess of 20 percent be undertaken, but the actual average returns are considerably smaller. Firms know that to attain the required return, they have to set high reservation levels. Our analysis also shows, as one would expect, that firms in an uncoordinated polyarchy do not take into account the negative externality that they exert on one another (i.e., each firm worsens the portfolio which the other firm faces) as much as they would were their reservation levels coordinated.

Comparative Statics of Reservation Levels. An immediate implication of (11)–(13) is the following.

PROPOSITION 5: *An unambiguous increase in the relative proportion of bad projects increases the reservation levels under both hierarchy and polyarchy. That is, if* $\partial g(x,\beta)/\partial\beta \lesseqgtr 0$ *as* $x \gtrless 0$, *then* $dR^s/d\beta > 0$ *for* $s = P, H,$ *and* C.

Thus, a worsening of the portfolio of available projects leads to tighter screening (lower probabilities of acceptance).

Another critical determinant of the reservation levels is the quality of information based on which projects are selected. Intuitively, we would expect a worsening of the quality of information to lead to higher reservation levels. For simplicity, let the noise associated with observing a project depend on x. Now consider a new distribution of θ, which first-order stochastically dominates the original distribution for $x < 0$. Because with the new distribution, we are more likely, at any reservation level, to accept a bad project, we refer to the new information as noisier than the original. It is easy to show

PROPOSITION 6: *If the screening is tight, noisier information induces an increase in the reservation level of a hierarchy.*[6]

Comparison of Performance. Because the reservation levels are set differently in the different systems, the *individuals'* Type-I and Type-II errors are different; this makes the comparison of the system performance more difficult than that in the previous section. We present three sets of results, focusing on the special case where there are two types of projects.

(a) Suppose a polyarchy has a larger output than a hierarchy when both systems use the hierarchy's (optimal) reservation levels. Then clearly, a *coordinated* polyarchy, using its reservation levels, will perform even better. Moreover, if Y^P is locally a concave function of R,[7] then Proposition 4 implies that an uncoordinated polyarchy using its own reservation levels, will also outperform hierarchy.

Thus, *when* $a = 1$, *and the reservation levels for hierarchy are such that the screening probabilities fall within the area ODA in Figure 4, then the output of a polyarchy is larger.* By the same reasoning, whenever the reservation levels for polyarchy are such that the screening probabilities fall within the area ADB, then a hierarchy has a larger output. Analogous interpretations apply when $a \neq 1$. Note that one cannot reach a verdict on the relative performance (using this approach) if hierarchy's screening probabilities fall within ADB, or if polyarchy's fall within ODA.

When will hierarchy's screening probabilities be such that at those probabilities, polyarchy outperforms hierarchy? First, consider the case where a is slightly less than unity. Then, $p_1 = p_2 = 0$ if we observe a completely uninformative signal. By continuity, using the above results, *if the signal concerning the quality of the project is sufficiently bad, then polyarchy outperforms hierarchy.* This is a

[6]This definition of an increase in noise is not the standard one in the statistical decision literature. Under the stated conditions, for all values of $x < 0$, $p(x, R)$ is increased for all R, while $p_R(xR)$ is decreased at the optimal R, provided only that the error density function for large θ is increased. This result establishes that each local maximum to Y^H is shifted to the right. To ensure that the global maximum is reduced, we need to assume either concavity, or to impose somewhat stronger conditions on how the distribution of noise changes.

[7]That is, $Y^P_{RR} < 0$ within the region $R^H < R < R^C$.

somewhat surprising result: one might have thought that with poor information, the second screening provided by hierarchy would be more valuable. But the reservation levels adjust so much, the resulting screening is so tight, that the second chance provided by polyarchy is more important than the second review provided by hierarchy.

Next, note from Proposition 5 that an improvement in the portfolio (i.e., an increase in a) leads to larger acceptance probabilities. Our earlier analysis, on the other hand, showed that a larger a increases the range of p_1 and p_2 within which polyarchy dominates a hierarchy. We therefore ascertain conditions under which, nonetheless, it can be established that polyarchy performs better than a hierarchy. For this, the internal optimum in a hierarchy, (11), is restated in the present case (where the initial portfolio consists of only two types of projects) as

$$ap_1 p_{1R} = p_2 p_{2R}. \tag{14}$$

Now define $k \equiv [m(R^H - z_1)/M(R^H - z_1)]/[m(R^H + z_2)/M(R^H + z_2)]$. Using (4), then, the following can easily be established.

PROPOSITION 7: *If* $k \leq 1$, *polyarchy performs better than hierarchy.*

To see what is entailed, consider the case where $m(\theta)$ is unimodal and a is large, so that reservation levels are low, sufficiently low that $R + z_2$ and $R - z_1$ are below the mode. Then $k < 1$ provided that for $R - z_1 < \theta < R + z_2$, $m(\theta)/M(\theta)$ is increasing in θ (i.e., $m_\theta > m^2/M$). Analogous results can be derived illustrating conditions under which at polyarchy's optimally chosen screening probabilities, hierarchy outperforms polyarchy.

(b) We can derive an alternative, somewhat weaker set of sufficient conditions under which one or the other system performs better by taking into account the fact that screening is tighter in a polyarchy. Let a polyarchy choose its reservation level such that the (expected) number of projects it undertakes is the same as that chosen optimally by a hierarchy. This is clearly not optimal, but if we can show that with this nonoptimal reservation level, polyarchy performs better than hierarchy, then a coordinated polyarchy will surely perform better with reservation levels optimally chosen.

If R^T denotes the reservation level at which a polyarchy chooses the same number of projects as that chosen by a hierarchy using reservation level R^H, then $R^T > R^H$. For brevity, we also define $m^P(x) \equiv m(R^T - x)$, and $m^H(x) \equiv m(R^H - x)$; M^P and M^H are defined analogously. Now, the above polyarchy performs better than hierarchy if the aggregate screening function is more discriminating for the former; that is, if $f_x^P > f_x^H$ in the relevant range[8] (equivalently, if $m^P M^P > m^H(1 - M^H)$). This will be true if the screening is very tight (in which case M^H is close to one), provided only that the difference between m^P and m^H is not too large, which, in turn, will be true provided m_θ is not too large. Moreover, with very tight screening, $R^P \approx R^C$ (i.e., the externality effect becomes negligible). It follows therefore that (under the conditions stated above) a hierarchy is outperformed by a coordinated as well as an uncoordinated polyarchy.

Much weaker conditions are required to establish the above result if R^T is not much larger than R^H. A sufficient condition in this case is that the screening is moderately tight, that is, a screen's probability of accepting a project is less than one-half.[9]

(c) We have investigated in detail the case of uniform distributions of errors, with mean zero, for the case of symmetric projects ($z_1 = z_2$). In this case: *A polyarchy has a*

[8] The probability that a project is accepted by a polyarchy is $f^P(R^T, x) = 1 - M^2(R^T - x)$, where $x = z_1$ and $-z_2$ for good and bad projects, respectively. The corresponding probability in a hierarchy is $f^H(R^H, x) = \{1 - M(R^H - x)\}^2$. With the same number of projects being accepted in the two systems, polyarchy performs better if $f^P(R^T, z_1) > f^H(R^H, z_1)$, or, if $f^P(R^T, -z_2) < (f^H, -z_2)$. An equivalent condition is $f^P(R^T, z_1) - f^P(R^T, -z_2) > f^H(R^H, z_1) - f^H(R^H, -z_2)$, which is satisfied if $f_x^P > f_x^H$ in the relevant range.

[9] This is because if $R^T \approx R^H$ then the required inequality (see fn. 8), $f^P(R^T, z_1) - f^P(R^T, -z_2) > f^H(R^H, z_1) - f^H(R^H, -z_2)$, is satisfied provided M is larger than $1/2$, within the range $R^H - z_1$ to $R^T + z_2$.

higher (lower) profit than a hierarchy if the proportion of good projects in the initial portfolio is less (more) than one-half. Obviously, if one hypothesizes that unprofitable ideas typically outnumber the profitable ones in a portfolio, then the present example suggests that a polyarchy is a superior institutional arrangement.

Remarks:

(*i*) The assumption of limited communication plays a critical role in our analysis. We assumed that firms in a polyarchy cannot (do not) communicate at all, and the bureaus in a hierarchy communicate only binary signals (whether they think a project is good or bad); they cannot communicate their actual information concerning the characteristics of the projects. We believe that although the extent of information sharing varies under different circumstances, it is seldom perfect, and our model has been constructed to capture the consequences of this.

(*ii*) The architecture of the economic system itself conveys some information to its constituents, which they use in setting decision rules. For example, in our analysis of polyarchy, in which the firms do not share any information with one another, each firm knows that some of the projects it receives are those rejected by the other firm and, consequently, the portfolio of projects faced by a firm is not an exact replica of the initial portfolio, but has been modified by the other firm. This implicit information is partly used in determing optimal reservation levels.

IV. Extensions

The basic components of our model are the screening function, the distribution of available projects, and the system's architecture, with its associated decision rule. Each of these components can be generalized (as we have partly done in our referenced papers). In the preceding section, we endogenized the screening function, given the available information. By allocating more resources to information acquisition, more informative signals can be obtained. The level of spending on information acquisition within each level of the hierarchy (by each firm within a polyarchy) must then be endogenously determined. In general, the levels of expenditure at each level of the hierarchy will not be the same: a higher quality screening at the lower bureau improves the portfolio to be evaluated by the higher bureau, but it also costs more because a larger number of projects are evaluated at the lower level. Similar issues arise in the assignment of individuals of differing abilities within a hierarchy (in the present paper all individuals have the same ability). The architecture of a system may also influence the mix of available projects because the likelihood of acceptance for projects of various types may well affect research incentives.

Alternative "architectures" that we have investigated include committees; that is, groups of individuals of different sizes who use particular decision rules (for instance, majority voting) to approve projects. (The decision-making units investigated in the present paper can be viewed as the limiting case of committees of one.) These polar architectures, in turn, can be viewed as building blocks for complex organizations and economies which are mixtures of hierarchies, polyarchies, and committees.

We can also investigate the consequences of alternative decision rules. We have followed the natural presumption that a project is not undertaken unless it is approved by the organization; within a hierarchy, by both bureaus; within a polyarchy, by at least one firm. We could, of course, *imagine* quite different organization of decision making: for instance, a hierarchy in which all projects are accepted except those which get vetoed by both bureaus and a polyarchy in which a project is accepted unless vetoed by one of the units.

We refer to organizations operating according to the veto rule as a veto hierarchy and polyarchy, in contrast to those we have analyzed earlier, which we refer to as an acceptance hierarchy and polyarchy. It can be easily shown that, in the absence of costs of coordination: *An acceptance polyarchy (hierarchy) is equivalent to a veto hierarchy (polyarchy)*.

Note, however, that the coordination requirements may be markedly different de-

pending on the architecture and the nature of decision rules. For instance, an acceptance polyarchy does not require any informational coordination among its constituent units; a firm does not need to inform other firms concerning the projects it has accepted or rejected. This is an important aspect of the independence of firms within market-like systems, which we stressed earlier. In comparison, in a veto polyarchy in which one unit can veto a project from being undertaken, each unit must inform other units which projects it has rejected. Similarly, in a veto hierarchy the lower bureau must send all projects to the higher bureau. If there are significant costs to informational coordination (significant noise in information transmission), then it is clear that an acceptance polyarchy may have an advantage over other organizational forms.

V. Applications and Conclusions

Our analysis has focused on alternative ways of structuring decision making that might be applied to any organization. The comparative statics propositions indicate the "objective" circumstances under which each form might be observed, if the choice of organizational form is being made explicitly or implicitly. Not only is such a comparative view relevant to corporate decision making, but also to decision making within the public sector: for instance, to on-going controversies over the organization of the military (three branches, more or less independent, or one unified armed force). Similarly, alternative political structures (with their systems of checks and balances) can be viewed as alternative architectures to balance the consequences of different types of human errors.

An important application of our approach, as we indicated in the beginning, is in the comparison of alternative ways of organizing economic systems, the polyarchical structure capturing certain central elements of market economies, the hierarchical structure those of the more centralized economies. The organization of decision making (and the corresponding errors, costs, and consequences) has played little or no role in some of the most important previous work in this area. There is clearly more at stake in the choice of an economic system than, for instance, a comparison of alternative algorithms for arriving at a once and for all allocation of society's resources, that was emphasized in the Lange-Lerner-Taylor claim of equivalence between price-guided socialist economies and market economies.

There are many aspects of the comparison of alternative systems with which we have not dealt adequately here, but which we believe can be incorporated into our framework, and which we hope to address in the future: for instance, the view that better incentive mechanisms can often be designed within decentralized economic systems (see Barry Nalebuff and Stiglitz, 1983), and the claim that centralized economic systems provide a better framework for dealing with externalities (which we have explicitly excluded from the analysis);[10] the view that a decentralized system's performance is less sensitive to the quality of the key decision makers; the view that natural selection mechanisms work more effectively in decentralized economics, and the view that decentralized structures provide greater stimulation for innovation.

Our analysis has, however, cast light on several other aspects of the debate concerning the relative merits of polyarchies vs. hierarchies: advocates of polyarchies point out that a good project has many opportunities of being accepted in their system, whereas critics contend that polyarchies fail to provide adequate checks against incompetent decision making. Critics of heirarchical structures claim that there are high costs to providing these checks; there are direct costs of additional evaluations, and there are indirect costs because good projects get rejected in the process of ensuring that bad projects do not get undertaken. Advocates of polyarchies point further to its virtues in

[10] In addition, we have assumed that only one firm can undertake a project. The inefficiencies which arise when many firms undertake similar projects within decentralized systems have been a source of criticism leveled at these systems.

economies of communication. There is a grain of truth in each of these views. In this paper, we have provided a framework within which one can assess the circumstances under which the grain of truth in one view is greater than that in the other.

REFERENCES

Lange, Oskar and Taylor, Fred M., *On the Economic Theory of Socialism*, New York: McGraw-Hill, 1964.

Nalebuff, Barry and Stiglitz, Joseph E., "Information, Competition and Markets," *American Economic Review Proceedings*, May 1983, *73*, 278–83.

Sah, Raaj Kumar and Stiglitz, Joseph E., (1985a) "The Architecture of Economic Systems: Hierarchies and Polyarchies," Economic Growth Center Discussion paper No. 476, Yale University, 1985.

______ **and** ______, (1985b) "Perpetuation, Self-Reproduction, and Evolution of Organizations: The Selection and Performance of Managers," paper presented at the 1985 World Congress of the Econometric Society, Cambridge, August 1985.

______ **and** ______, (1985c) "Economics of Committees," Economic Growth Center Discussion Paper No. 486, Yale University, 1985.

Fairness as a Constraint on Profit Seeking: Entitlements in the Market

By DANIEL KAHNEMAN, JACK L. KNETSCH, AND RICHARD THALER*

Community standards of fairness for the setting of prices and wages were elicited by telephone surveys. In customer or labor markets, it is acceptable for a firm to raise prices (or cut wages) when profits are threatened and to maintain prices when costs diminish. It is unfair to exploit shifts in demand by raising prices or cutting wages. Several market anomalies are explained by assuming that these standards of fairness influence the behavior of firms.

Just as it is often useful to neglect friction in elementary mechanics, there may be good reasons to assume that firms seek their maximal profit as if they were subject only to legal and budgetary constraints. However, the patterns of sluggish or incomplete adjustment often observed in markets suggest that some additional constraints are operative. Several authors have used a notion of fairness to explain why many employers do not cut wages during periods of high unemployment (George Akerlof, 1979; Robert Solow, 1980). Arthur Okun (1981) went further in arguing that fairness also alters the outcomes in what he called customer markets—characterized by suppliers who are perceived as making their own pricing decisions, have some monopoly power (if only because search is costly), and often have repeat business with their clientele. Like labor markets, customer markets also sometimes fail to clear:

> ...firms in the sports and entertainment industries offer their customers tickets at standard prices for events that clearly generate excess demand. Popular new models of automobiles may have waiting lists that extend for months. Similarly, manufacturers in a number of industries operate with backlogs in booms and allocate shipments when they obviously could raise prices and reduce the queue. [p. 170]

Okun explained these observations by the hostile reaction of customers to price increases that are not justified by increased costs and are therefore viewed as unfair. He also noted that customers appear willing to accept "fair" price increases even when demand is slack, and commented that "...in practice, observed pricing behavior is a vast distance from do it yourself auctioneering" (p. 170).

The argument used by these authors to account for apparent deviations from the simple model of a profit-maximizing firm is that fair behavior is instrumental to the maximization of long-run profits. In Okun's model, customers who suspect that a supplier treats them unfairly are likely to start searching for alternatives; Akerlof (1980, 1982) suggested that firms invest in their reputation to produce goodwill among their customers and high morale among their employees; and Arrow argued that trusted suppliers may be able to operate in markets that are otherwise devastated by the lemons problem (Akerlof, 1970; Kenneth Arrow, 1973). In these approaches, the rules of fairness define the terms of an enforceable im-

*Kahneman: Department of Psychology, University of California, Berkeley, CA 94720; Knetsch, Department of Economics, Simon Fraser University; Thaler: Johnson School of Management, Cornell University. The research was carried out when Kahneman was at the University of British Columbia. It was supported by the Department of Fisheries and Oceans Canada. Kahneman and Thaler were also supported by the U.S. Office of Naval Research and the Alfred P. Sloan Foundation, respectively. Conversations with J. Brander, R. Frank and A. Tversky were very helpful.

plicit contract: Firms that behave unfairly are punished in the long run. A more radical assumption is that some firms apply fair policies even in situations that preclude enforcement—this is the view of the lay public, as shown in a later section of this paper.

If considerations of fairness do restrict the actions of profit-seeking firms, economic models might be enriched by a more detailed analysis of this constraint. Specifically, the rules that govern public perceptions of fairness should identify situations in which some firms will fail to exploit apparent opportunities to increase their profits. Near-rationality theory (Akerlof and Janet Yellen, 1985) suggests that such failures to maximize by a significant number of firms in a market can have large aggregate effects even in the presence of other firms that seek to take advantage of all available opportunities. Rules of fairness can also have significant economic effects through the medium of regulation. Indeed, Edward Zajac (forthcoming) has inferred general rules of fairness from public reactions to the behavior of regulated utilities.

The present research uses household surveys of public opinions to infer rules of fairness for conduct in the market from evaluations of particular actions by hypothetical firms.[1] The study has two main objectives: (*i*) to identify community standards of fairness that apply to price, rent, and wage setting by firms in varied circumstances; and (*ii*) to consider the possible implications of the rules of fairness for market outcomes.

The study was concerned with scenarios in which a *firm* (merchant, landlord, or employer) makes a pricing or wage-setting decision that affects the outcomes of one or more *transactors* (customers, tenants, or employees). The scenario was read to the participants, who evaluated the fairness of the action as in the following example:

Question 1. A hardware store has been selling snow shovels for $15. The morning after a large snowstorm, the store raises the price to $20. Please rate this action as:

Completely Fair Acceptable
Unfair Very Unfair

The two favorable and the two unfavorable categories are grouped in this report to indicate the proportions of respondents who judged the action acceptable or unfair. In this example, 82 percent of respondents ($N = 107$) considered it unfair for the hardware store to take advantage of the short-run increase in demand associated with a blizzard.

The approach of the present study is purely descriptive. Normative status is not claimed for the generalizations that are described as "rules of fairness," and the phrase "it is fair" is simply an abbreviation for "a substantial majority of the population studied thinks it fair." The paper considers in turn three determinants of fairness judgments: the reference transaction, the outcomes to the firm and to the transactors, and the occasion for the action of the firm. The final sections are concerned with the enforcement of fairness and with economic phenomena that the rules of fairness may help explain.

I. Reference Transactions

A central concept in analyzing the fairness of actions in which a firm sets the terms of future exchanges is the *reference transaction*, a relevant precedent that is characterized by a reference price or wage, and by a positive reference profit to the firm. The treatment is restricted to cases in which the fairness of the reference transaction is not itself in question.

The main findings of this research can be summarized by a principle of *dual entitlement*, which governs community standards of fairness: Transactors have an entitlement to the terms of the reference transaction and firms are entitled to their reference profit. A firm is not allowed to increase its profits by

[1]Data were collected between May 1984 and July 1985 in telephone surveys of randomly selected residents of two Canadian metropolitan areas: Toronto and Vancouver. Equal numbers of adult female and male respondents were interviewed for about ten minutes in calls made during evening hours. No more than five questions concerned with fairness were included in any interview, and contrasting questions that were to be compared were never put to the same respondents.

arbitrarily violating the entitlement of its transactors to the reference price, rent or wage (Max Bazerman, 1985; Zajac, forthcoming). When the reference profit of a firm is threatened, however, it may set new terms that protect its profit at transactors' expense.

Market prices, posted prices, and the history of previous transactions between a firm and a transactor can serve as reference transactions. When there is a history of similar transactions between firm and transactor, the most recent price, wage, or rent will be adopted for reference unless the terms of the previous transaction were explicitly temporary. For new transactions, prevailing competitive prices or wages provide the natural reference. The role of prior history in wage transactions is illustrated by the following pair of questions:

Question 2A. A small photocopying shop has one employee who has worked in the shop for six months and earns $9 per hour. Business continues to be satisfactory, but a factory in the area has closed and unemployment has increased. Other small shops have now hired reliable workers at $7 an hour to perform jobs similar to those done by the photocopy shop employee. The owner of the photocopying shop reduces the employee's wage to $7.

($N = 98$) Acceptable 17% Unfair 83%

Question 2B. A small photocopying shop has one employee...[as in Question 2A]...The current employee leaves, and the owner decides to pay a replacement $7 an hour.

($N = 125$) Acceptable 73% Unfair 27%

The current wage of an employee serves as reference for evaluating the fairness of future adjustments of that employee's wage—but not necessarily for evaluating the fairness of the wage paid to a replacement. The new worker does not have an entitlement to the former worker's wage rate. As the following question shows, the entitlement of an employee to a reference wage does not carry over to a new labor transaction, even with the same employer:

Question 3. A house painter employs two assistants and pays them $9 per hour. The painter decides to quit house painting and go into the business of providing landscape services, where the going wage is lower. He reduces the workers' wages to $7 per hour for the landscaping work.

($N = 94$) Acceptable 63% Unfair 37%

Note that the same reduction in wages that is judged acceptable by most respondents in Question 3 was judged unfair by 83 percent of the respondents to Question 2A.

Parallel results were obtained in questions concerning residential tenancy. As in the case of wages, many respondents apply different rules to a new tenant and to a tenant renewing a lease. A rent increase that is judged fair for a new lease may be unfair for a renewal. However, the circumstances under which the rules of fairness require landlords to bear such opportunity costs are narrowly defined. Few respondents consider it unfair for the landlord to sell the accommodation to another landlord who intends to raise the rents of sitting tenants, and even fewer believe that a landlord should make price concessions in selling an accommodation to its occupant.

The relevant reference transaction is not always unique. Disagreements about fairness are most likely to arise when alternative reference transactions can be invoked, each leading to a different assessment of the participants' outcomes. Agreement on general principles of fairness therefore does not preclude disputes about specific cases (see also Zajac, forthcoming). When competitors change their price or wage, for example, the current terms set by the firm and the new terms set by competitors define alternative reference transactions. Some people will consider it unfair for a firm not to raise its wages when competitors are increasing theirs. On the other hand, price increases that are not justified by increasing costs are judged less objectionable when competitors have led the way.

It should perhaps be emphasized that the reference transaction provides a basis for fairness judgments because it is normal, not necessarily because it is just. Psychological studies of adaptation suggest that any stable state of affairs tends to become accepted eventually, at least in the sense that alterna-

tives to it no longer readily come to mind. Terms of exchange that are initially seen as unfair may in time acquire the status of a reference transaction. Thus, the gap between the behavior that people consider fair and the behavior that they expect in the marketplace tends to be rather small. This was confirmed in several scenarios, where different samples of respondents answered the two questions: "What does fairness require?" and "What do you think the firm would do?" The similarity of the answers suggests that people expect a substantial level of conformity to community standards—and also that they adapt their views of fairness to the norms of actual behavior.

II. The Coding of Outcomes

It is a commonplace that the fairness of an action depends in large part on the signs of its outcomes for the agent and for the individuals affected by it. The cardinal rule of fair behavior is surely that one person should not achieve a gain by simply imposing an equivalent loss on another.

In the present framework, the outcomes to the firm and to its transactors are defined as gains and losses in relation to the reference transaction. The transactor's outcome is simply the difference between the new terms set by the firm and the reference price, rent, or wage. The outcome to the firm is evaluated with respect to the reference profit, and incorporates the effect of exogenous shocks (for example, changes in wholesale prices) which alter the profit of the firm on a transaction at the reference terms. According to these definitions, the outcomes in the snow shovel example of Question 1 were a $5 gain to the firm and a $5 loss to the representative customer. However, had the same price increase been induced by a $5 increase in the wholesale price of snow shovels, the outcome to the firm would have been nil.

The issue of how to define relevant outcomes takes a similar form in studies of individuals' preferences and of judgments of fairness. In both domains, a descriptive analysis of people's judgments and choices involves rules of *naive accounting* that diverge in major ways from the standards of rationality assumed in economic analysis. People commonly evaluate outcomes as gains or losses relative to a neutral reference point rather than as endstates (Kahneman and Amos Tversky, 1979). In violation of normative standards, they are more sensitive to out-of-pocket costs than to opportunity costs and more sensitive to losses than to foregone gains (Kahneman and Tversky, 1984; Thaler, 1980). These characteristics of evaluation make preferences vulnerable to framing effects, in which inconsequential variations in the presentation of a choice problem affect the decision (Tversky and Kahneman, 1986).

The entitlements of firms and transactors induce similar asymmetries between gains and losses in fairness judgments. An action by a firm is more likely to be judged unfair if it causes a loss to its transactor than if it cancels or reduces a possible gain. Similarly, an action by a firm is more likely to be judged unfair if it achieves a gain to the firm than if it averts a loss. Different standards are applied to actions that are elicited by the threat of losses or by an opportunity to improve on a positive reference profit —a psychologically important distinction which is usually not represented in economic analysis.

Judgments of fairness are also susceptible to framing effects, in which form appears to overwhelm substance. One of these framing effects will be recognized as the money illusion, illustrated in the following questions:

Question 4A. A company is making a small profit. It is located in a community experiencing a recession with substantial unemployment but no inflation. There are many workers anxious to work at the company. The company decides to decrease wages and salaries 7% this year.

($N = 125$) Acceptable 38% Unfair 62%

Question 4B. ...with substantial unemployment and inflation of 12%...The company decides to increase salaries only 5% this year.

($N = 129$) Acceptable 78% Unfair 22%

Although the real income change is approximately the same in the two problems, the judgments of fairness are strikingly different. A wage cut is coded as a loss and consequently judged unfair. A nominal raise

which does not compensate for inflation is more acceptable because it is coded as a gain to the employee, relative to the reference wage.

Analyses of individual choice suggest that the disutility associated with an outcome that is coded as a loss may be greater than the disutility of the same objective outcome when coded as the elimination of a gain. Thus, there may be less resistance to the cancellation of a discount or bonus than to an equivalent price increase or wage cut. As illustrated by the following questions, the same rule applies as well to fairness judgments.

Question 5A. A shortage has developed for a popular model of automobile, and customers must now wait two months for delivery. A dealer has been selling these cars at list price. Now the dealer prices this model at $200 above list price.

($N=130$) Acceptable 29% Unfair 71%

Question 5B. ...A dealer has been selling these cars at a discount of $200 below list price. Now the dealer sells this model only at list price.

($N=123$) Acceptable 58% Unfair 42%

The significant difference between the responses to Questions 5A and 5B (*chi-squared* = 20.91) indicates that the $200 price increase is not treated identically in the two problems. In Question 5A the increase is clearly coded as a loss relative to the unambiguous reference provided by the list price. In Question 5B the reference price is ambiguous, and the change can be coded either as a loss (if the reference price is the discounted price), or as the elimination of a gain (if the reference price is the list price). The relative leniency of judgments in Question 5B suggests that at least some respondents adopted the latter frame. The following questions illustrate the same effect in the case of wages:

Question 6A. A small company employs several people. The workers' incomes have been about average for the community. In recent months, business for the company has not increased as it had before. The owners reduce the workers' wages by 10 percent for the next year.

($N=100$) Acceptable 39% Unfair 61%

Question 6B. A small company employs several people. The workers have been receiving a 10 percent annual bonus each year and their total incomes have been about average for the community. In recent months, business for the company has not increased as it had before. The owners eliminate the workers' bonus for the year.

($N=98$) Acceptable 80% Unfair 20%

III. Occasions for Pricing Decisions

This section examines the rules of fairness that apply to three classes of occasions in which a firm may reconsider the terms that it sets for exchanges. (*i*) *Profit reductions*, for example, by rising costs or decreased demand for the product of the firm. (*ii*) *Profit increases*, for example, by efficiency gains or reduced costs. (*iii*) *Increases in market power*, for example, by temporary excess demand for goods, accommodations or jobs.

A. *Protecting Profit*

A random sample of adults contains many more customers, tenants, and employees than merchants, landlords, or employers. Nevertheless, most participants in the surveys clearly consider the firm to be entitled to its reference profit: They would allow a firm threatened by a reduction of its profit below a positive reference level to pass on the entire loss to its transactors, without compromising or sharing the pain. By large majorities, respondents endorsed the fairness of passing on increases in wholesale costs, in operating costs, and in the costs associated with a rental accommodation. The following two questions illustrate the range of situations to which this rule was found to apply.

Question 7. Suppose that, due to a transportation mixup, there is a local shortage of lettuce and the wholesale price has increased. A local grocer has bought the usual quantity of lettuce at a price that is 30 cents

per head higher than normal. The grocer raises the price of lettuce to customers by 30 cents per head.

($N=101$) Acceptable 79% Unfair 21%

Question 8. A landlord owns and rents out a single small house to a tenant who is living on a fixed income. A higher rent would mean the tenant would have to move. Other small rental houses are available. The landlord's costs have increased substantially over the past year and the landlord raises the rent to cover the cost increases when the tenant's lease is due for renewal.

($N=151$) Acceptable 75% Unfair 25%

The answers to the last question, in particular, indicate that it is acceptable for firms to protect themselves from losses even when their transactors suffer substantial inconvenience as a result. The rules of fairness that yield such judgments do not correspond to norms of charity and do not reflect distributional concerns.

The attitude that permits the firm to protect a positive reference profit at the transactors' expense applies to employers as well as to merchants and landlords. When the profit of the employer in the labor transaction falls below the reference level, reductions of even nominal wages become acceptable. The next questions illustrate the strong effect of this variable.

Question 9A. A small company employs several workers and has been paying them average wages. There is severe unemployment in the area and the company could easily replace its current employees with good workers at a lower wage. The company has been making money. The owners reduce the current workers' wages by 5 percent.

($N=195$) Acceptable 23% Unfair 77%

Question 9B. ...The company has been losing money. The owners reduce the current workers' wages by 5 percent.

($N=195$) Acceptable 68% Unfair 32%

The effect of firm profitability was studied in greater detail in the context of a scenario in which Mr. Green, a gardener who employs two workers at $7 an hour, learns that other equally competent workers are willing to do the same work for $6 an hour. Some respondents were told that Mr. Green's business was doing well, others were told that it was doing poorly. The questions, presented in open format, required respondents to state "what is fair for Mr. Green to do in this situation," or "what is your best guess about what Mr. Green would do...." The information about the current state of the business had a large effect. Replacing the employees or bargaining with them to achieve a lower wage was mentioned as fair by 67 percent of respondents when business was said to be poor, but only by 25 percent of respondents when business was good. The proportion guessing that Mr. Green would try to reduce his labor costs was 75 percent when he was said to be doing poorly, and 49 percent when he was said to be doing well. The differences were statistically reliable in both cases.

A firm is only allowed to protect itself at the transactor's expense against losses that pertain directly to the transaction at hand. Thus, it is unfair for a landlord to raise the rent on an accommodation to make up for the loss of another source of income. On the other hand, 62 percent of the respondents considered it acceptable for a landlord to charge a higher rent for apartments in one of two otherwise identical buildings, because a more costly foundation had been required in the construction of that building.

The assignment of costs to specific goods explains why it is generally unfair to raise the price of old stock when the price of new stock increases:

Question 10. A grocery store has several months supply of peanut butter in stock which it has on the shelves and in the storeroom. The owner hears that the wholesale price of peanut butter has increased and immediately raises the price on the current stock of peanut butter.

($N=147$) Acceptable 21% Unfair 79%

The principles of naive accounting apparently include a FIFO method of inventory cost allocation.

B. *The Allocation of Gains*

The data of the preceding section could be interpreted as evidence for a cost-plus rule of fair pricing, in which the supplier is expected to act as a broker in passing on marked-up costs (Okun). A critical test of this possible rule arises when the supplier's costs diminish: A strict cost-plus rule would require prices to come down accordingly. In contrast, a dual-entitlement view suggests that the firm is only prohibited from increasing its profit by causing a loss to its transactors. Increasing profits by retaining cost reductions does not violate the transactors' entitlement and may therefore be acceptable.

The results of our previous study (1986) indicated that community standards of fairness do not in fact restrict firms to the reference profit when their costs diminish, as a cost-plus rule would require. The questions used in these surveys presented a scenario of a monopolist supplier of a particular kind of table, who faces a $20 reduction of costs on tables that have been selling for $150. The respondents were asked to indicate whether "fairness requires" the supplier to lower the price, and if so, by how much. About one-half of the survey respondents felt that it was acceptable for the supplier to retain the entire benefit, and less than one-third would require the supplier to reduce the price by $20, as a cost-plus rule dictates. Further, and somewhat surprisingly, judgments of fairness did not reliably discriminate between primary producers and middlemen, or between savings due to lower input prices and to improved efficiency.

The conclusion that the rules of fairness permit the seller to keep part or all of any cost reduction was confirmed with the simpler method employed in the present study.

Question 11A. A small factory produces tables and sells all that it can make at $200 each. Because of changes in the price of materials, the cost of making each table has recently decreased by $40. The factory reduces its price for the tables by $20.

($N = 102$) Acceptable 79% Unfair 21%

Question 11B. ...the cost of making each table has recently decreased by $20. The factory does not change its price for the tables.

($N = 100$) Acceptable 53% Unfair 47%

The even division of opinions on Question 11B confirms the observations of the previous study. In conjunction with the results of the previous section, the findings support a dual-entitlement view: the rules of fairness permit a firm not to share in the losses that it imposes on its transactors, without imposing on it an unequivocal duty to share its gains with them.

C. *Exploitation of Increased Market Power*

The market power of a firm reflects the advantage to the transactor of the exchange which the firm offers, compared to the transactor's second-best alternative. For example, a blizzard increases the surplus associated with the purchase of a snow shovel at the regular price, compared to the alternatives of buying elsewhere or doing without a shovel. The respondents consider it unfair for the hardware store to capture any part of the increased surplus, because such an action would violate the customer's entitlement to the reference price. Similarly, it is unfair for a firm to exploit an excess in the supply of labor to cut wages (Question 2A), because this would violate the entitlement of employees to their reference wage.

As shown by the following routine example, the opposition to exploitation of shortages is not restricted to such extreme circumstances:

Question 12. A severe shortage of Red Delicious apples has developed in a community and none of the grocery stores or produce markets have any of this type of apple on their shelves. Other varieties of apples are plentiful in all of the stores. One grocer receives a single shipment of Red Delicious apples at the regular wholesale cost and raises the retail price of these Red Delicious apples by 25% over the regular price.

($N = 102$) Acceptable 37% Unfair 63%

Raising prices in response to a shortage is unfair even when close substitutes are read-

ily available. A similar aversion to price rationing held as well for luxury items. For example, a majority of respondents thought it unfair for a popular restaurant to impose a $5 surcharge for Saturday night reservations.

Conventional economic analyses assume as a matter of course that excess demand for a good creates an opportunity for suppliers to raise prices, and that such increases will indeed occur. The profit-seeking adjustments that clear the market are in this view as natural as water finding its level—and as ethically neutral. The lay public does not share this indifference. Community standards of fairness effectively require the firm to absorb an opportunity cost in the presence of excess demand, by charging less than the clearing price or paying more than the clearing wage.

As might be expected from this analysis, it is unfair for a firm to take advantage of an increase in its monopoly power. Respondents were nearly unanimous in condemning a store that raises prices when its sole competitor in a community is temporarily forced to close. As shown in the next question, even a rather mild exploitation of monopoly power is considered unfair.

Question 13. A grocery chain has stores in many communities. Most of them face competition from other groceries. In one community the chain has no competition. Although its costs and volume of sales are the same there as elsewhere, the chain sets prices that average 5 percent higher than in other communities.

($N=101$) Acceptable 24% Unfair 76%

Responses to this and two additional versions of this question specifying average price increases of 10 and 15 percent did not differ significantly. The respondents clearly viewed such pricing practices as unfair, but were insensitive to the extent of the unwarranted increase.

A monopolist might attempt to increase profits by charging different customers as much as they are willing to pay. In conventional theory, the constraints that prevent a monopolist from using perfect price discrimination to capture all the consumers' surplus are asymmetric information and difficulties in preventing resale. The survey results suggest the addition of a further restraint: some forms of price discrimination are outrageous.

Question 14. A landlord rents out a small house. When the lease is due for renewal, the landlord learns that the tenant has taken a job very close to the house and is therefore unlikely to move. The landlord raises the rent $40 per month more than he was planning to do.

($N=157$) Acceptable 9% Unfair 91%

The near unanimity of responses to this and similar questions indicates that an action that deliberately exploits the special dependence of a particular individual is exceptionally offensive.

The introduction of an explicit auction to allocate scarce goods or jobs would also enable the firm to gain at the expense of its transactors, and is consequently judged unfair.

Question 15. A store has been sold out of the popular Cabbage Patch dolls for a month. A week before Christmas a single doll is discovered in a storeroom. The managers know that many customers would like to buy the doll. They announce over the store's public address system that the doll will be sold by auction to the customer who offers to pay the most.

($N=101$) Acceptable 26% Unfair 74%

Question 16. A business in a community with high unemployment needs to hire a new computer operator. Four candidates are judged to be completely qualified for the job. The manager asks the candidates to state the lowest salary they would be willing to accept, and then hires the one who demands the lowest salary.

($N=154$) Acceptable 36% Unfair 64%

The auction is opposed in both cases, presumably because the competition among potential buyers or employees benefits the firm. The opposition can in some cases be mitigated by eliminating this benefit. For example, a sentence added to Question 15, indicating that "the proceeds will go to

UNICEF" reduced the negative judgments of the doll auction from 74 to 21 percent.

The strong aversion to price rationing in these examples clearly does not extend to all uses of auctions. The individual who sells securities at twice the price paid for them a month ago is an object of admiration and envy—and is certainly not thought to be gouging. Why is it fair to sell a painting or a house at the market-clearing price, but not an apple, dinner reservation, job, or football game ticket? The rule of acceptability appears to be this: Goods for which an active resale market exists, and especially goods that serve as a store of value, can be sold freely by auction or other mechanisms allowing the seller to capture the maximum price. When resale is a realistic possibility, which is not the case for most consumer goods, the potential resale price reflects the higher value of the asset and the purchaser is therefore not perceived as sustaining a loss.

IV. Enforcement

Several considerations may deter a firm from violating community standards of fairness. First, a history or reputation of unfair dealing may induce potential transactors to take their business elsewhere, because of the element of trust that is present in many transactions. Second, transactors may avoid exchanges with offending firms at some cost to themselves, even when trust is not an issue. Finally, the individuals who make decisions on behalf of firms may have a preference for acting fairly. The role of reputation effects is widely recognized. This section presents some indications that a willingness to resist and to punish unfairness and an intrinsic motivation to be fair could also contribute to fair behavior in the marketplace.

A willingness to pay to resist and to punish unfairness has been demonstrated in incentive compatible laboratory experiments. In the ultimatum game devised by Werner Guth, Rolf Schmittberger, and Bernd Schwarze (1982), the participants are designated as allocators or recipients. Each allocator anonymously proposes a division of a fixed amount of money between himself (herself) and a recipient. The recipient either accepts the offer or rejects it, in which case both players get nothing. The standard game theoretic solution is for the allocator to make a token offer and for the recipient to accept it, but Guth et al. observed that many allocators offer an equal division and that recipients sometimes turn down positive offers. In our more detailed study of resistance to unfairness (1986), recipients were asked to indicate in advance how they wished to respond to a range of possible allocations: A majority of participants were willing to forsake $2 rather than accept an unfair allocation of $10.

Willingness to punish unfair actors was observed in another experiment, in which subjects were given the opportunity to share a sum of money evenly with one of two anonymous strangers, identified only by the allocation they had proposed to someone else in a previous round. About three-quarters of the undergraduate participants in this experiment elected to share $10 evenly with a stranger who had been fair to someone else, when the alternative was to share $12 evenly with an unfair allocator (see our other paper).

A willingness to punish unfairness was also expressed in the telephone surveys. For example, 68 percent of respondents said they would switch their patronage to a drugstore five minutes further away if the one closer to them raised its prices when a competitor was temporarily forced to close; and, in a separate sample, 69 percent indicated they would switch if the more convenient store discriminated against its older workers.

The costs of enforcing fairness are small in these examples—but effective enforcement in the marketplace can often be achieved at little cost to transactors. Retailers will have a substantial incentive to behave fairly if a large number of customers are prepared to drive an extra five minutes to avoid doing business with an unfair firm. The threat of future punishment when competitors enter may also deter a temporary monopolist from fully exploiting short-term profit opportunities.

In traditional economic theory, compliance with contracts depends on enforcement. It is a mild embarrassment to the standard model that experimental studies often pro-

duce fair behavior even in the absence of enforcement (Elizabeth Hoffman and Matthew Spitzer, 1982, 1985; our paper, 1986; Arvin Roth, Michael Malouf, and J. Keith Murninghan, 1981; Reinhard Selten, 1978). These observations, however, merely confirm common sense views of human behavior. Survey results indicate a belief that unenforced compliance to the rules of fairness is common. This belief was examined in two contexts: tipping in restaurants and sharp practice in automobile repairs.

Question 17A. If the service is satisfactory, how much of a tip do you think most people leave after ordering a meal costing $10 in a restaurant that they visit frequently?

($N=122$) Mean response = $1.28

Question 17B. ...in a restaurant on a trip to another city that they do not expect to visit again?

($N=124$) Mean response = $1.27

The respondents evidently do not treat the possibility of enforcement as a significant factor in the control of tipping. Their opinion is consistent with the widely observed adherence to a 15 percent tipping rule even by one-time customers who pay and tip by credit card, and have little reason to fear embarrassing retaliation by an irate server.

The common belief that tipping is controlled by intrinsic motivation can be accommodated with a standard microeconomic model by extending the utility function of individuals to include guilt and self-esteem. A more difficult question is whether firms, which the theory assumes to maximize profits, also fail to exploit some economic opportunities because of unenforced compliance with rules of fairness. The following questions elicited expectations about the behavior of a garage mechanic dealing with a regular customer or with a tourist.

Question 18A. [A man leaves his car with the mechanic at his regular / A tourist leaves his car at a] service station with instructions to replace an expensive part. After the [customer/tourist] leaves, the mechanic examines the car and discovers that it is not necessary to replace the part; it can be repaired cheaply. The mechanic would make much more money by replacing the part than by repairing it. Assuming the [customer/tourist] cannot be reached, what do you think the mechanic would do in this situation?

Make more money by replacing the part
customer: 60% tourist: 63%

Save the customer money by repairing the part
customer: 40% tourist: 37%

Question 18B. Of ten mechanics dealing with a [regular customer/tourist], how many would you expect to save the customer money by repairing the part?

Mean response
customer: 3.62 tourist: 3.72

The respondents do not approach garages with wide-eyed naive faith. It is therefore all the more noteworthy that they expect a tourist and a regular customer to be treated alike, in spite of the obvious difference between the two cases in the potential for any kind of enforcement, including reputation effects.[2]

Here again, there is no evidence that the public considers enforcement a significant factor. The respondents believe that most mechanics (usually excluding their own) would be less than saintly in this situation. However, they also appear to believe that the substantial minority of mechanics who would treat their customers fairly are not motivated in each case by the anticipation of sanctions.

V. Economic Consequences

The findings of this study suggest that many actions that are both profitable in the short run and not obviously dishonest are likely to be perceived as unfair exploitations of market power.[3] Such perceptions can have

[2] Other respondents were asked to assess the probable behavior of their own garage under similar circumstances: 88 percent expressed a belief that their garage would act fairly toward a regular customer, and 86 percent stated that their garage would treat a tourist and a regular customer similarly.

[3] This conclusion probably holds in social and cultural groups other than the Canadian urban samples

significant consequences if they find expression in legislation or regulation (Zajac, 1978; forthcoming). Further, even in the absence of government intervention, the actions of firms that wish to avoid a reputation for unfairness will depart in significant ways from the standard model of economic behavior. The survey results suggest four propositions about the effects of fairness considerations on the behavior of firms in customer markets, and a parallel set of hypotheses about labor markets.

A. *Fairness in Customer Markets*

PROPOSITION 1: *When excess demand in a customer market is unaccompanied by increases in suppliers' costs, the market will fail to clear in the short run.*

Evidence supporting this proposition was described by Phillip Cagan (1979), who concluded from a review of the behavior of prices that, "Empirical studies have long found that short-run shifts in demand have small and often insignificant effects [on prices]" (p. 18). Other consistent evidence comes from studies of disasters, where prices are often maintained at their reference levels although supplies are short (Douglas Dacy and Howard Kunreuther, 1969).

A particularly well-documented illustration of the behavior predicted in proposition 1 is provided by Alan Olmstead and Paul Rhode (1985). During the spring and summer of 1920 there was a severe gasoline shortage in the U.S. West Coast where Standard Oil of California (SOCal) was the dominant supplier. There were no government-imposed price controls, nor was there any threat of such controls, yet SOCal reacted by imposing allocation and rationing schemes while maintaining prices. Prices were actually higher in the East in the absence of any shortage. Significantly, Olmstead and Rhode note that the eastern firms had to purchase crude at higher prices while SOCal, being vertically integrated, had no such excuse for raising price. They conclude from confidential SOCal documents that SOCal officers "...were clearly concerned with their public image and tried to maintain the appearance of being 'fair'" (p. 1053).

PROPOSITION 2: *When a single supplier provides a family of goods for which there is differential demand without corresponding variation of input costs, shortages of the most valued items will occur.*

There is considerable support for this proposition in the pricing of sport and entertainment events, which are characterized by marked variation of demand for goods or services for which costs are about the same (Thaler, 1985). The survey responses suggest that charging the market-clearing price for the most popular goods would be judged unfair.

Proposition 2 applies to cases such as those of resort hotels that have in-season and out-of-season rates which correspond to predictable variations of demand. To the extent that constraints of fairness are operating, the price adjustments should be insufficient, with excess demand at the peak. Because naive accounting does not properly distinguish between marginal and average costs, customers and other observers are likely to adopt off-peak prices as a reference in evaluating the fairness of the price charged to peak customers. A revenue-maximizing (low) price in the off-season may suggest that the profits achievable at the peak are unfairly high. In spite of a substantial degree of within-season price variation in resort and ski hotels, it appears to be the rule that most of these establishments face excess demand during the peak weeks. One industry explanation is: "If you gouge them at Christmas, they won't be back in March."

PROPOSITION 3: *Price changes will be more responsive to variations of costs than to variations of demand, and more responsive to cost increases than to cost decreases.*

The high sensitivity of prices to short-run variations of costs is well documented

studied here, although the detailed rules of fairness for economic transactions may vary.

(Cagan). The idea of asymmetric price rigidity has a history of controversy (Timur Kuran, 1983; Solow; George Stigler and James Kindahl, 1970) and the issue is still unsettled. Changes of currency values offer a potential test of the hypothesis that cost increases tend to be passed on quickly and completely, whereas cost decreases can be retained at least in part. When the rate of exchange between two currencies changes after a prolonged period of stability, the prediction from Proposition 3 is that upward adjustments of import prices in one country will occur faster than the downward adjustments expected in the other.

PROPOSITION 4: *Price decreases will often take the form of discounts rather than reductions in the list or posted price.*

This proposition is strongly supported by the data of Stigler and Kindahl. Casual observation confirms that temporary discounts are much more common than temporary surcharges. Discounts have the important advantage that their subsequent cancellation will elicit less resistance than an increase in posted price. A temporary surcharge is especially aversive because it does not have the prospect of becoming a reference price, and can only be coded as a loss.

B. *Fairness in Labor Markets*

A consistent finding of this study is the similarity of the rules of fairness that apply to prices, rents, and wages. The correspondence extends to the economic predictions that may be derived for the behavior of wages in labor markets and of prices in customer markets. The first proposition about prices asserted that resistance to the exploitation of short-term fluctuations of demand could prevent markets from clearing. The corresponding prediction for labor markets is that wages will be relatively insensitive to excess supply.

The existence of wage stickiness is not in doubt, and numerous explanations have been offered for it. An entitlement model of this effect invokes an implicit contract between the worker and the firm. Like other implicit contract theories, such a model predicts that wage changes in a firm will be more sensitive to recent firm profits than to local labor market conditions. However, unlike the implicit contract theories that emphasize risk shifting (Costas Azariadis, 1975; Martin Baily, 1974; Donald Gordon, 1974), explanations in terms of fairness (Akerlof, 1979, 1982; Okun; Solow) lead to predictions of wage stickiness even in occupations that offer no prospects for long-term employment and therefore provide little protection from risk. Okun noted that "Casual empiricism about the casual labor market suggests that the Keynesian wage floor nonetheless operates; the pay of car washers or stock clerks is seldom cut in a recession, even when it is well above any statutory minimum wage" (1981, p. 82), and he concluded that the employment relation is governed by an "invisible handshake," rather than by the invisible hand (p. 89).

The dual-entitlement model differs from a Keynesian model of sticky wages, in which nominal wage changes are always nonnegative. The survey findings suggest that nominal wage cuts by a firm that is losing money or threatened with bankruptcy do not violate community standards of fairness. This modification of the sticky nominal wage dictum is related to Proposition 3 for customer markets. Just as they may raise prices to do so, firms may also cut wages to protect a positive reference profit.

Proposition 2 for customer markets asserted that the dispersion of prices for similar goods that cost the same to produce but differ in demand will be insufficient to clear the market. An analogous case in the labor market involves positions that are similar in nominal duties but are occupied by individuals who have different values in the employment market. The prediction is that differences in income will be insufficient to eliminate the excess demand for the individuals considered most valuable, and the excess supply of those considered most dispensable. This prediction applies both within and among occupations.

Robert Frank (1985) found that the individuals in a university who already are the most highly paid in each department are also

the most likely targets for raiding. Frank explains the observed behavior in terms of envy and status. An analysis of this phenomenon in terms of fairness is the same as for the seasonal pricing of resort rooms: Just as prices that clear the market at peak demand will be perceived as gouging if the resort can also afford to operate at off-peak rates, a firm that can afford to pay its most valuable employees their market value may appear to grossly underpay their less-valued colleagues. A related prediction is that variations among departments will also be insufficient to clear the market. Although salaries are higher in academic departments that compete with the private sector than in others, the ratio of job openings to applicants is still lower in classics than in accounting.

The present analysis also suggests that firms that frame a portion of their compensation package as bonuses or profit sharing will encounter relatively little resistance to reductions of compensation during slack periods. This is the equivalent of Proposition 4. The relevant psychological principle is that losses are more aversive than objectively equivalent foregone gains. The same mechanism, combined with the money illusion, supports another prediction: Adjustments of real wages will be substantially greater in inflationary periods than in periods of stable prices, because the adjustments can then be achieved without making nominal cuts—which are always perceived as losses and are therefore strongly resisted. An unequal distribution of gains is more likely to appear fair than a reallocation in which there are losers.

This discussion has illustrated several ways in which the informal entitlements of customers or employees to the terms of reference transactions could enter an economic analysis. In cases such as the pricing of resort facilities, the concern of customers for fair pricing may permanently prevent the market from clearing. In other situations, the reluctance of firms to impose terms that can be perceived as unfair acts as a friction-like factor. The process of reaching equilibrium can be slowed down if no firm wants to be seen as a leader in moving to exploit changing market conditions. In some instances an initially unfair practice (for example, charging above list price for a popular car model) may spread slowly until it evolves into a new norm—and is no longer unfair. In all these cases, perceptions of transactors' entitlements affect the substantive outcomes of exchanges, altering or preventing the equilibria predicted by an analysis that omits fairness as a factor. In addition, considerations of fairness can affect the form rather than the substance of price or wage setting. Judgments of fairness are susceptible to substantial framing effects, and the present study gives reason to believe that firms have an incentive to frame the terms of exchanges so as to make them appear "fair."

REFERENCES

Akerlof, George A., "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, August 1970, *84*, 488–500.

______, "The Case against Conservative Macroeconomics: An Inaugural Lecture," *Economica*, August 1979, *46*, 219–37.

______, "A Theory of Social Custom, of Which Unemployment May Be One Consequence," *Quarterly Journal of Economics*, June 1980, *94*, 749–75.

______, "Labor Contracts as Partial Gift Exchange," *Quarterly Journal of Economics*, November 1982, *97*, 543–69.

______ **and Yellen, Janet L.**, "Can Small Deviations from Rationality Make Significant Differences to Economic Equilibrium?," *American Economic Review*, September 1985, *75*, 708–20.

Arrow, Kenneth, "Social Responsibility and Economic Efficiency," *Public Policy*, Summer 1973, *21*, 303–17.

Azariadis, Costas, "Implicit Contracts and Unemployment Equilibria," *Journal of Political Economy*, December 1975, *83*, 1183–202.

Baily, Martin N., "Wages and Employment Under Uncertain Demand," *Review of Economic Studies*, January 1974, *41*, 37–50.

Bazerman, Max H., "Norms of Distributive Justice in Interest Arbitration," *Industrial

and Labor Relations Review, July 1985, *38*, 558–70.

Cagan, Phillip, *Persistent Inflation: Historical and Policy Essays*, New York: Columbia University Press, 1979.

Dacy, Douglas C. and Kunreuther, Howard, *The Economics of Natural Disasters*, New York: Free Press, 1969.

Frank, Robert H., *Choosing the Right Pond; Human Behavior and the Quest for Status*, New York: Oxford University Press, 1985.

Gordon, Donald F., "A Neo-Classical Theory of Keynesian Unemployment," *Economic Inquiry*, December 1974, *12*, 431–59.

Guth, Werner, Schmittberger, Rolf and Schwarze, Bernd, "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization*, 1982, *3*, 367–88.

Hoffman, Elizabeth and Spitzer, Matthew L., "The Coase Theorem: Some Experimental Tests," *Journal of Law and Economics*, April 1982, *25*, 73–98.

______ **and** ______ "Entitlements, Rights, and Fairness: An Experimental Examination of Subjects' Concepts of Distributive Justice," *Journal of Legal Studies*, June 1985, *14*, 259–97.

Kahneman, Daniel, Knetsch, Jack L. and Thaler, Richard, "Fairness and the Assumptions of Economics," *Journal of Business*, forthcoming, 1986.

______ **and Tversky, Amos,** "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, March 1979, *47*, 263–91.

______ **and** ______, "Choices, Values, and Frames," *American Psychologist*, April 1984, *39*, 341–50.

Kuran, Timur, "Asymmetric Price Rigidity and Inflationary Bias," *American Economic Review*, June 1983, *73*, 373–82.

Okun, Arthur, *Prices and Quantities: A Macroeconomic Analysis*, Washington: The Brookings Institution, 1981.

Olmstead, Alan L. and Rhode, Paul, "Rationing Without Government: The West Coast Gas Famine of 1920," *American Economic Review*, December 1985, *75*, 1044–55.

Roth, Alvin, Malouf, Michael and Murnighan, J. Keith, "Sociological Versus Strategic Factors in Bargaining," *Journal of Economic Behavior and Organization*, June 1981, *2*, 153–77.

Selten, Reinhard, "The Equity Principle in Economic Behavior," in Hans W. Gottinger and Werner Leinfellner, eds., *Decision Theory and Social Ethics: Issues in Social Choice*, Dordrecht: D. Reidel, 1978, 289–301.

Solow, Robert M., "On Theories of Unemployment," *American Economic Review*, March 1980, *70*, 1–11.

Stigler, George J. and Kindahl, James K., *The Behavior of Industrial Prices*, NBER, New York: Columbia University Press, 1970.

Thaler, Richard, "Toward a Positive Theory of Consumer Choice," *Journal of Economic Behavior and Organization*, March 1980, *1*, 39–60.

______, "Mental Accounting and Consumer Choice," *Marketing Science*, Summer 1985, *4*, 199–214.

Tversky, Amos and Kahneman, Daniel, "Rational Choice and the Framing of Decisions," *Journal of Business*, forthcoming, 1986.

Zajac, Edward E., *Fairness or Efficiency: An Introduction to Public Utility Pricing*, Cambridge: Ballinger, 1978.

______, "Perceived Economic Justice: The example of Public Utility Regulation," in H. Peyton Young, ed., *Cost Allocation: Methods, Principles and Applications*, Amsterdam: North-Holland, forthcoming.

Measuring the Spillovers from Technical Advance: Mainframe Computers in Financial Services

By TIMOTHY F. BRESNAHAN*

Measuring the social gains from recent technological advances is difficult because there are no real output indexes for some important adopters. Measurement methods that infer the willingness to pay of the adopting industries from the derived demand curve for a new technology overcome this difficulty. The derived demand for high-speed computers for use in banks, finance, and insurance is shown to imply a very large social gain to computerization that was not captured by manufacturers of computers.

An important component of the social benefits of technical advances is the spillover to the customers of the advancing sectors. Process innovations lower costs and therefore prices, to the benefit of downstream sectors. New or improved products whose prices do not fully reflect their enhanced downstream value yield a similar spillover. Measurement of the quantitative importance of such spillovers[1] contributes to our understanding of the role of technical advance in long-term economic growth,[2] and is central to the formation of public policies towards innovation.[3] As Zvi Griliches (1979) points out, important postwar technological advances, such as those in electronics and health, have largely benefited downstream sectors in which the spillovers are hard to measure. The downstream sectors—services, government, health care, etc.—lack sensible measures of real output, so that calculation of the impact of the new technology on productivity or cost is difficult. The goal of this paper is to devise methods for the measurement of spillovers which do not depend on the existence of real output indexes in the downstream sector. Instead, the value spilled over will be inferred from the demand curve of the downstream sector for the output of the advancing sector.

If an innovation serves only to lower the price of a consumption good, the resulting area under the demand curve for the good (adjusted for income effects) is a welfare index: calculating it yields the value of the spillover. In the more usual case of technological progress in an intermediate goods industry, the welfare economics are more complicated. The results of this paper show that the area under the *derived demand curve* for an intermediate input (adjusted for income effects) is a welfare index as well. The derived demand curve is the relationship between the price and quantity of an input, allowing for equilibrium in the output market. Along the derived demand curve, input quantity demanded will increase in response to a fall in input price both because of substitution out of other factors and because a lower price for one input will usually

*Assistant Professor of Economics, Stanford University, Stanford, CA 94305. Research on this paper was supported by National Science Foundation grant IST-8507536 to the Stanford Center for Economic Policy Research. W. Edward Steinmueller and two referees made extensive comments on an earlier draft. I am indebted to Paul David, John Pencavel, and Bronwyn Hall for many useful conversations, but remain responsible for any errors.

[1]Another kind of "spillover" is also discussed in the economics of technology—the external benefits of knowledge creation. When investment in new knowledge by one firm "spills over" to other firms, a positive externality arises within the advancing sector itself. For an attempt to quantify that other kind of spillovers, see Adam Jaffe (1984).

[2]See Robert Fogel (1964) and Paul David (1975) for discussions of the economics of major innovations and for discussion of the importance of the railroads in nineteenth-century U.S. economic growth.

[3]See Edwin Mansfield et al. (1977) for a discussion of the welfare economics of spillovers. They also measure the spillovers from a number of recent technological advances.

lead to an expansion of output.[4] If price in the upstream sector falls, the area under the demand curve for its output measures the sum of the increased producer's surplus in the downstream sectors plus the consumer's surplus of final demanders.

The real quality-adjusted price of computers has fallen rapidly since they first became commercially available thirty years ago, largely because of technological advances in the high-speed computer sector itself and advances in electronic components such as integrated circuits. Further, a very large fraction of computer sales is to sectors such as services and government in which real output is poorly measured. The national income accounting convention for these sectors has been that real outputs are measured by inputs, and research has not yet provided convincing alternative indexes of real output. As a result, total factor productivity (*TFP*) indexes for these sectors measure an accounting identity rather than the output and productivity effects of the technical change in computing. Inferring the value of the technology from adopters' willingness to pay is an attractive, feasible alternative.

This paper will calculate the spillovers from advances in general purpose ("mainframe") computers to the financial services sector (banking, insurance, and brokerage) from 1958 to 1972. This sector made up a large fraction of private demand for computers. The indexes calculated will measure welfare under two central assumptions. First, existing quality-adjusted price indexes for computers are reliable. Second, the financial services sector is competitive on those margins not fixed by regulation. Competition will lead the sector to act as the agent of its customers. Thus we can treat purchases of computers by the services sector as if they were made for those customers, and infer the gain to the customers from service's derived demand for computers.

The area under the demand curve for a product can be measured either by econometric techniques or by index-number techniques. (This is the analog of the distinction between econometric production function estimation and the calculation of an index of *TFP* growth.) In this paper, index-number techniques based closely on those of Douglas Caves, Laurits Christensen, and Erwin Diewert (1982) will be used. These techniques measure the real resource cost of providing consumers with a given level of utility, or the real costs of a given level of output in producer sectors which use financial services. These costs were lowered substantially by falls in the price of computers as inputs to the financial services sector.

I. The Welfare Analysis of Derived Demand

The theoretical model used in the spillover calculation has three elements. The first is an upstream sector (computers in the application reported below) which produces a vector of outputs Z_1 at quality-adjusted prices W_1. Both Z_1 and W_1 are observable to the analyst. It is useful to think of W_1 as falling rapidly over time because of technical progress.[5]

The second element in the model is the downstream sector (banking, finance and insurance) which produces outputs Q_1 using Z. The vector Z includes all of Z_1 as elements. The prices of all inputs are denoted W and assumed to be observable to the analyst. Each output in Q_1 has an endogenous price and quality; the vectors of these are called P_1 and Δ_1. Both P_1 and Δ_1 are unobservable to the analyst. This unobservability arises because product quality in this sector is difficult to measure quantitatively. Only expenditures on each good, j, in the sector $p_j \times q_j$, are observable. The technology in this sector is represented by a con-

[4]This is Alfred Marshall's usage: the demand for the input is "derived" from that for the output (*Principles of Economics*, 1948, p. 381). The alternative usage of "derived demand" to mean "demand for an input holding the price of the output fixed" is incorrect and should be discontinued.

[5]If there is substantial market power in the Z_1 sector, prices will fall more slowly than they otherwise might, and downstream sectors will reap smaller benefits from technical advance in any given period. This paper will only quantify the benefits actually reaped by the downstream sector from the actual price fall, not the potential benefits.

stant returns to scale production function for each element of Q_1:

$$q_j = F_j\left(\delta_j, Z^j\right), \tag{1}$$

where δ_j is output quality, and Z^j the inputs used in this particular industry. Price and quality setting in the Q_1 sector are competitive, except that the government regulates prices in some industries in this sector (for good or ill). For convenience, let the regulated industries be the first R elements of Q_1. Let g_j stand for the government-set price in industry j, and let G_R and Q_R refer to the vectors of prices and quantities in the R regulated industries. The rule by which government gets G_R will be discussed below. For firms under regulation, the behavioral assumption is that they still act competitively with respect to the price, quality, or marketing variables they do control.

The third element of the model is made up by the consumers and the firms in other industries who are customers of the Q_1 sector. Both kinds of customers have demands which depend on the quality and the price of outputs from the Q_1 sector. Both kinds of customers buy non-Q_1 goods as well; the prices of these are taken to be observed.

This section has two purposes: to show that the unobservability of Q_1, P_1, and Δ_1 is of no consequence in attempting to measure the social value of the fall in W_1, and to derive index-number formulae which yield an estimate of that value. For brevity, I explicitly treat only the case in which all of the output of the Q_1 sector is purchased by consumers.[6]

A. *Derived Demand by Competitive Industries*

The cost function for industry j is denoted $C_j(q_j, \delta_j, W)$. Consumers buy not only Q_1 but also other goods Q_2 at prices P_2. These prices are taken to be observable. Consumers have utility function $u = U(Q_1, \Delta_1, Q_2)$, and associated expenditure function $E(u, P_2, P_1, \Delta_1)$. The downstream industries are competitive with supply functions:

$$\begin{aligned} (2)\qquad & p_j = g_j, \qquad j \in R \\ & p_j = \frac{\partial C_j(q_j, \delta_j, W)}{\partial q_j}, \qquad \text{otherwise,} \\ & -\frac{\partial E}{\partial \delta_j} = \frac{\partial C_j(q_j, \delta_j, W)}{\partial \delta_j} \qquad \forall j \end{aligned}$$

Let M_R be an R vector and $X'Y$ be the inner product of vectors X and Y. Then we have the result that market equilibrium will act so as to minimize the sum of social costs plus (marginal) regulatory costs:

Result 1. *If consumers are expenditure minimizers and firms in the Q_1 sector are competitive cost minimizers, then Z, Δ_1 and Q solve*:

$$\min E = P_2'Q_2 + \sum_{j=1}^{J} W'Z^j + (G_R - M_R)'Q_R \qquad **$$

$$\textit{subject to} \quad u = U(Q_1, \Delta_1, Q_2) \qquad *1$$

$$\textit{subject to} \quad q_j = F_j\left(\delta_j, Z^j\right) \quad j = 1, \ldots, J \qquad *2$$

when M_j is evaluated at marginal cost in regulated industry j.

PROOF:

The proof, in the Appendix, shows that the first-order conditions for $**$, taken together with the definitions of cost and expenditure functions, are the same as (2).

To understand the intuition and the utility of Result 1, first consider the case in which there are no regulated industries in the Q_1 sector. Under competition, each p_j will maximize the sum of producer's and consumer's surplus (ignoring income effects for brevity) in market j. As a result, we can treat these industries as if they were owned by consumers. Calculating a welfare index on the basis of Result 1 will involve using P_2 and

[6]An earlier, longer version of this paper has a fuller treatment. See also fn. 27.

W, since P_1 does not appear. This calculation is based on the prices of inputs into the Q_1 sector instead of outputs: it is the calculation of a cost of living and of making Q index. From this viewpoint, there is no more difficulty with the lack of a consistent price index for insurance services than with a lack of a consistent price index for refrigeration services. In the latter case, I calculate the real cost-of-living index based on the inputs the household purchases in order to make refrigeration services, such as refrigerators and electricity. In the former case, we can calculate an index using the inputs bought by the insurance industry "for" households.

More formally, let $E^*(u, P_2, W)$ be the solution to $**$ in the case where there are no regulated industries. It is immediate from the usual envelope theorem argument that

$$z_k = \sum_j z_k^j = \frac{\partial E^*(u, P_2, W)}{\partial w_k}.$$

As a result, integration of the derived demand functions Z will yield estimates of E^*. This calculation is taken up below.

This argument cannot be applied to the regulated industry case without alteration. Result 1 is stated as if $G_R - M_R$ is exogenous, since competitive price-regulated firms act as if they cannot affect these margins. But quality competition renders the level of marginal cost endogenous, and even the government-set prices of regulated industries must surely be treated as responsive to costs in the long run. Obviously, no general results can be obtained in the regulated industry case, but two important models of regulated industry price setting have similar implications.

First, consider the model of government behavior in which prices in the long run are ρ_j percent greater than marginal costs in industry j. Let $\tilde{E}(u, P_2, W, G_R - M_R)$ be the solution to $**$ in general. Differentiation of $\tilde{E}$ does not determine all of the equilibrium values since $G_R - M_R$ is endogenous. The observable derived demand curves will take the form $Z^*(u, P_2, W) = \tilde{Z}(u, P_2, W, G_R(P_2, W) - M_R(P_2, W))$. The regulations imply

$$g_j - m_j = \rho_j \frac{\partial C_j(\tilde{q}_j, \tilde{\delta}_j, W)}{\partial q_j}$$

where the tildes denote equilibrium values. By how much will an increase in w_k lower social surplus?[7]

$$d\tilde{E}/dw_k = \sum_j z_k^j$$

$$+ \sum_j \rho_j Q_j \left[\frac{\partial^2 C_j}{\partial q_j^2} \frac{\partial \tilde{q}_j}{\partial z_k} + \frac{\partial^2 C_j}{\partial q_j \partial \delta_j} \frac{\partial \tilde{\delta}_j}{\partial z_k} \right].$$

If the equilibrium effect of higher factor prices is to raise marginal costs, the term in square brackets will be positive. As a result, integration of the demand curves Z^* will understate the welfare effects of changes in W.

This result appears to be in conflict with the observation that regulated-industry quality competition can lead to too rapid adoption of technologies whose prices are falling over time (Nathan Rosenberg and David Mowery, 1982). The conflict is only apparent. Quality competition may in fact lead to too much demand for some inputs Z, relative to the unregulated first-best. However, we are seeking an index of surplus actually achieved in the market, not of surplus that could be achieved if the regulations were removed.

The second model of regulation to consider is that of successful cartelization of the industry through regulation. Then the industry equilibrium, including the equation

[7]It is not quite right to treat $\tilde{E}$ as social costs at the margin, since the excess of price over marginal costs in the regulated industry is presumably a return to someone and therefore part of social surplus. Thus the calculation made here tends to understate social surplus. If quality competition destroys most or all of the regulatory rents, this distinction will be unimportant. (This need not always follow from quality competition: see George Stigler, 1968, and Lawrence White, 1972.)

for the regulatory price, acts to maximize profit, not social surplus. The derived demands Z will once again provide an index of what is being maximized in the market: profit, not total surplus. However, the surplus actually obtained in the market is larger than profit, since consumers do obtain some surplus. Therefore welfare indexes based on derived demand by a monopolized industry will tend to understate the true social value of upstream price changes.[8]

When certain prices are not easily observable, an opportunity to substitute theory for data arises; it is possible to use assumptions about the process by which those prices are set to replace information about the prices themselves. When the prices are set in a competitive market, it is appropriate to do exactly what one would expect: infer value from the area under the derived demand curve. When the market is not competitive, this argument fails. However, the same calculation does yield underestimates of the social gain from a decline in upstream prices.

B. *Implied Welfare Measures*

The result of the last subsection can be exploited in deriving welfare measures. Suppose we have two different time periods, 0 and 1, and wish to know how much the fall in W_1 from W_1^0 to W_1^1 benefited consumers.[9] The first task is to figure out by how much more consumers are better off in period 1 than in period 0. Answers to this question take the form of cost-of-living indexes: how much more expensive would it have been to provide period 1 utility in period 0, and how much cheaper would it have been to provide period 0 utility in period 1? The primary difference of the current model from the usual analysis of a cost-of-living index is that here I analyze the cost of living and of making Q_1, since the model (implicitly) treats the Q_1 sector as if it were owned by consumers. The relative cost of providing period 1 utility at period 0 prices is

$$C^{*1}=\frac{E^*(u^1,P_2^0,W^0)}{E^*(u^1,P_2^1,W^1)}.$$

The inverse of the relative cost of providing period 0 utility at period 1 prices is

$$C^{*0}=\frac{E^*(u^0,P_2^0,W^0)}{E^*(u^0,P_2^1,W^1)}.$$

The large body of useful results from the theory of index numbers for the standard consumer problem apply here without alteration. The derived demand curves from the market equilibrium, $Z(u,P_2,W)$, are the same ones that would come from directly maximizing the problem $**$. As a result, the marginal social value of elements of Z is revealed in the derived demand in exactly the same sense that a single consumer's demands reveal the marginal value of goods to that consumer. Any functional form assumptions about $E(\ )$ lead to function forms for the cost-of-living index which can be calculated without econometric estimation. In the regulated case, these indexes will be biased against assigning welfare gains to falls in regulated industry factor prices. I follow Caves et al. in the choice of functional form, and assume that E^* has the translog functional form in both periods 0 and 1, and further that any changes in technology or tastes between the two periods shift only the first-order terms of the translog function.[10] The practical impact of these assumptions will be discussed below.

Under these assumptions, an estimate of the real cost of living and of making Q_1 index can be calculated on the basis of the observables alone. Let E^t denote the value of E^* at time t. Let s_Z^t be the vector of shares of expenditures on Z in E^t. Similarly, s_Q^t is the vector of shares of the Q_2 consumer goods in E^t with typical elements $s_{Q,k}$. The elements of the vector (s_Z^t, s_Q^t) sum to one.

[8] This single-industry argument has been made more generally in Robert Willig's (1983) analysis of taxation in many industries with market power.

[9] In what follows, superscripts denote time periods.

[10] See Caves et al. for translog formulae.

Then (following Caves et al.),

$$(3)\quad \frac{1}{2}\log(C^{*0}\times C^{*1}) = \frac{1}{2}\left(\sum_{k=1}^{K}\left(s_{Q,k}^{0}+s_{Q,k}^{1}\right)\log\left(\frac{p_{2,k}^{0}}{p_{2,k}^{1}}\right) + \sum_{n=1}^{N}\left(s_{Z,n}^{0}+s_{Z,n}^{1}\right)\log\left(\frac{w_n^0}{w_n^1}\right)\right).$$

Suppose a cost-of-living index were calculated using incorrect price and quantity indexes $\hat{P}_1$ and $\hat{Q}_1$. (These might be the indexes available from the national accounts.) In the translog case, the percentage bias to such an index is[11]

$$(4)\quad \frac{1}{2}\left(\sum_{j=1}^{J}\left(\sigma_{\hat{Q}_j}^{0}+\sigma_{\hat{Q}_j}^{1}\right)\log\left(\frac{\hat{P}_j^0}{\hat{P}_j^1}\right) - \sum_{n=1}^{N}\left(s_{Z,n}^{0}+s_{Z,n}^{1}\right)\log\left(\frac{w_n^0}{w_n^1}\right)\right)$$

where $\sigma_{\hat{Q}_1}$ is the vector of Q_1 shares in consumer expenditure as normally defined. Calculation of this bias would reveal by how much growth in real consumption was understated because of the use of the incorrect price indexes for the Q_1 sector.

The second welfare analysis to undertake is the answer to a counterfactual question: by how much are the downstream sectors and their customers better off than if there had not been the extraordinary technical advance in the Z_1 sector? This requires some assessment of how high W_1 would have been in period 1 without the technical advance. This is clearly arbitrary in any application, but a useful baseline is to take the relative price of Z_1 in period 0 and hold it constant in the counterfactual period. Labeling the counterfactual period period 2, we have $P_2^2 = P_2^1$, $W_2^2 = W_2^1$, and $W_1^2 = I(P_2^1, W_2^1)/I(P_2^0, W_2^0)\times W_1^0$ where $I()$ is some price index formula (the translog in the application below).

If the demand for the output of an advancing sector has been estimated econometrically, measuring the gain to downstream sectors from a fall in the price-performance ratio is straightforward. One simply integrates the demand system to obtain the function E^* and directly calculates the cost of living index at the counterfactual prices W_1^2 (following Jerry Hausman, 1981). In performing this analysis, the econometrician runs the risk of out-of-sample extrapolation, since the counterfactual prices have never been observed. But if the econometric maintained hypothesis is still valid even at the counterfactual point, the resulting welfare estimates will be meaningful.

The index-number method used here requires a further assumption about what would have happened in the counterfactual no-technical-advance case. The percentage amount of spillovers is a cost of living and of making Q_1 index from the counterfactual period to period 1. As in equation (3), this will take the form:

$$(5)\quad \frac{1}{2}\left(\sum_{k=1}^{K}\left(s_{Q_{2,k}}^{2}+s_{Q_{2,k}}^{1}\right)\log\left(\frac{p_{2,k}^{2}}{p_{2,k}^{1}}\right) + \sum_{n=1}^{N}\left(s_{Z,n}^{2}+s_{Z,n}^{1}\right)\log\left(\frac{w_n^2}{w_n^1}\right)\right).$$

Since $W_1^1 = W_1^2$ and $P_2^1 = P_2^2$, (5) immediately reduces to

$$(6)\quad \frac{1}{2}\sum_{n=1}^{N1}\left(s_{Z,n}^{2}+s_{Z,n}^{1}\right)\log\left(\frac{w_n^2}{w_n^1}\right),$$

where $N1$ is the number of elements of Z_1. Note that (6) depends on the shares of Z_1 in the counterfactual period. Some method of estimating these shares must be undertaken in order to calculate the indexes. There is no best means by which to do this, so in the application I will report a sensitivity analysis with respect to s_Z^2.

The methods used here for calculating welfare gains to downstream sectors from

[11] This calculation assumes that the value of inputs equals the value of outputs in the Q_1 sector, as in national income accounting data.

upstream technical advance are familiar from single-person decision theory. Their extension here to market equilibrium, and to the case of endogenous product quality, has served primarily to clarify the accounting conventions. The greatest benefit of an explicit theory will be seen in the next section, where the appropriateness of the assumptions for the application can be rigorously assessed.

II. Mainframe Computers in the Financial Services Sector, 1958–72

In this section, Z_1 is general-purpose (mainframe) computers. The downstream sector into which computers spill over, Q_1, is the financial services sector (FSS), defined as banking, brokerage, insurance, and related businesses.[12] This is a natural application, since the FSS was an early adapter of computer technology, primarily for narrowly defined data processing activities such as bookkeeping and recordkeeping.[13] Firms in this sector variously purchased, rented, and leased mainframe computers, and purchased computer services from "service bureaus."[14] Many of the real output and price indexes used in the national accounts for industries in the FSS use an "inputs = outputs" approach.[15] I will make no attempt to construct better indexes.[16] Instead, the problem will be avoided by using the theory to treat the FSS as if it were owned by its customers.

The trick of treating a sector as the agent of its customers will be used twice. Because of the importance of service bureaus in providing computer services to the FSS, we will treat the service bureaus as if they were owned by the FSS. The fraction of the computers, other capital, labor, and intermediate inputs of the service bureaus that correspond to FSS purchases will be accounted as if they were bought by the FSS directly.[17] I will compare years for which an input-output (IO) table exists. The base year of 1958 had the first IO table following the commercial introduction of mainframe computers. Thus, period 0 is 1958.

Period 1 is set to 1972 for several reasons. Historically, few of the industries in the FSS had free-market price setting. Interest rates for the liabilities of financial intermediaries were set by various federal government bodies. Insurance prices were in part set by state regulatory bodies. Until 1975, brokerage fees on the New York Stock Exchange, were set by the SEC. The fact of regulation itself does not make our calculations irrelevant, as was shown above. A serious problem would arise in the use of the index-number methodology, however, in comparing a regulated with an unregulated period. For this reason, my calculations end in 1972 rather than with a later IO table, as the regulatory environment of the FSS was substantially liberalized over the 1970's. A second reason for stopping in 1972 is that extension of a quality-adjusted price index for computers past 1972 would be quite difficult. Over the 1970's, more and

[12] Financial services as defined here are sector 70 in the input/output tables of the Department of Commerce Bureau of Economic Analysis (BEA) and are SIC 60–64 and 67.

[13] See Montgomery Phister (1979, ch. 3, especially part 3.1). See Horst Brand and John Duke (1982) for an interesting review of labor-saving computer installations in banking.

[14] As late as 1975, 71 percent of the banks using computer services purchased them from an external service bureau. Many small banks purchase computer services from larger money-center banks with which they are in correspondent banking services relationships. Insurance companies and brokers tend to own or lease computers, not buy computer services. See Philip Wyborg et al. (1977, p. 22).

[15] See John Kendrick (1982b) for summary and useful discussion of BEA methods of estimating real output in the services sector generally. Within the FSS, the BEA estimates real output for banking, credit agencies other than banks, and security and commodity brokers and services by a weighted labor-hours indicator. Insurance brokers, agents and services have real outputs estimated by an index of transactions volume.

[16] Extension of the transactions volume approach to commercial banking has been proposed by John Gorman (1969). Jerome Mark (1982) and Brand and Duke report some estimates using this approach.

[17] Service bureaus come into existence for a variety of reasons. In this instance, the most important motivation is probably to spread lumpy computer capital over many users. Thus it is probably appropriate to draw no distinctions between input prices at service bureaus and in the FSS.

TABLE 1—EXPENDITURES AND SHARES OF BEA SECTOR 70

	Expenditures, Current $\$10^6$ (1)	Share in Direct Cost (2)	Share in Total Cost (3)	Share in Cost with BEA 73 Added (4)
1958				
Labor	8492	.405	.545	.574
GP Computers	34	.00159	.002150	.00225
Other Capital	793	.038	.051	.053
FSS (BEA 70)	5389	.257	–	–
Business Services (BEA 73)	1025	.049	.066	–
Other Interm.	5231	.250	.336	.370
1972				
Labor	28595	.422	.548	.597
GP Computers	1237	.0183	.0237	.0254
Other Capital	3959	.0584	.0759	.1050
FSS (BEA 70)	15562	.2298	–	–
Business Services (BEA 73)	6033	.0891	.1156	–
Other Interm.	12321	.1819	.2362	.2724

more of the tasks previously assigned to mainframe hardware were moved to software as a result of operating systems advances. At the same time, competition between small mainframes and larger minicomputers began to increase.[18]

A. *Share and Price Calculations for the FSS*

The first task is to construct factor shares and factor prices for the FSS in 1958 and 1972. The approach used here relies heavily on data developed by Frank Gollop and Dale Jorgenson (1980, 1983) Jack Faucett Associates (JFA) (1979, 1977), and Barbara Fraumeni and Jorgenson (1980). Table 1 shows expenditures and shares of the FSS in 1958 and 1972, first giving the gross (current) dollar expenditure on six factors of production. These differ from JFA in that the consumption by the FSS of its own output as an intermediate input, and the consumption of the output of business services by the FSS are treated separately.[19] The remaining columns show the cost shares of the remaining four factors of production after these two inputs are treated as if they were made in the FSS. The correction for purchases of own-input takes the form (using superscripts to indicate columns) $s_j^{(3)} = s_j^{(2)}/(1 - s_{70}^{(2)})$ for all inputs j other than the FSS (BEA 70) input. In column 4, inputs into the business services sector are distributed to the FSS. Since the output of the business services sector is treated as if it were produced in the FSS, it is necessary to allocate labor, intermediate inputs, and computer and other capital purchased by the business services industry to the FSS. Thus $s_j^{(4)} = s_j^{(3)} + s_{73}^{(3)} \times s_j^{73}$, where s_j^{73} is the share of input j into the business services sector.[20]

The capital input has been divided into two components, computer and noncomputer. Here "computer" means only the CPU and memory of a mainframe computer system. Expenditures on computers are calcu-

[18] Robert Gordon (1985) reports, however, a slowing in the rate of decline in quality-adjusted computer prices after 1972. I therefore speculate that extension of my results to the present would find slowing returns to computerization.

[19] These are calculated from the 1958 and 1972 *IO* tables, as reported in the *Survey of Current Business*, November 1964, February 1979, and April 1979. Service Bureaus are a subset of BEA 73, "Business Services."

[20] This procedure underestimates the share of computer services in the FSS. The factor shares used in this calculation are for the entire services sector. Firms in the FSS, however, tend to concentrate their business services purchases in the computer service firms more than the average firm in the economy.

lated as if the computers were rented, regardless of their actual ownership. Rental prices are list prices for different model computers as reported by Montgomery Phister (1979). The inventory of privately held computers by model and of computers held in banking and insurance in Phister is used to weight the rentals.[21]

The treatment of complements to computers such as software and computer peripherals (especially mass storage devices) deserves some comment. There was considerable technical progress in software and peripherals over the period studied, both caused by and enhancing computerization. In such circumstances, there is no unambiguous definition of "the benefits of cheaper computers," if this term is something that is added together with "the benefits of improved software and peripherals." Some of the increase in the quantity of computers demanded actually caused by technical advance in those other inputs may well be attributed to technical advance in computers in my calculations. On the other hand, falls in the price of computers led to increased demands for the complements, which is not counted as increases in value here.[22] My calculations underestimate the total value of computers plus peripherals in the FSS. One might further adopt the convention that the correct definition of the area under the demand curve for right shoes per se is one-half the area under the demand curve for shoes. On this convention, it is not clear whether I over- or underestimate the area under the demand curve for computers per se.

The small share of computers in 1958 FSS expenditures, just over two-tenths of a percent in column 4 in Table 1, naturally raises a question about the validity of the translog approximation to E^* over such a broad

Table 2—Price Indexes

	1972 = 1.00
Labor	.482
GP Computers	
Gordon	333.62
Chow	19.37
Other Capital	.704
Intermediate	.6607
FSS (BEA 70)	.5354

range. It would be very troubling if the behavior of the isoquants near the computers = 0 axis was driving the calculations. Consideration of (4) and (6) shows this concern to be unfounded. If the share of computers in 1958 were doubled or halved, the calculations would change little. This intuition is straightforward: roughly speaking, (6) calculates the size of a welfare triangle. The translog works better if not literally extended to the axes, but evaluated at some "small" value at which the demand curve "almost" cuts the axis. What determines the size of the triangle is not how small small is, but rather how high the price is at which the demand curve almost cuts the axis.

Price indexes for the FSS and for the inputs into it are reported in Table 2. The index for the FSS itself is the implicit deflator as calculated by the BLS.[23] The wage rate is taken directly from Gollop and Jorgenson (1983). The price index for remaining intermediate goods is calculated using the JFA (1977) procedure. The noncomputer capital service price is calculated using the JFA data following the procedure of Fraumeni and Jorgenson, except that real estate sector capital is excluded.

It is clear from the formulae in (4) and (6), and from the geometrical argument made above, that the welfare inferences drawn here will be quite sensitive to the quality-adjusted price index for computers. A large literature on the economic measurement of quality-adjusted prices for computers has followed on the work of Gregory Chow (1967).[24] This

[21] See Phister's Tables II.2.10, II.3.11.4, II.3.11.5. Expenditures on noncomputer capital are calculated by first following the JFA procedure and then subtracting out the computer expenditures.

[22] Both of these statements are contingent on the empirical assertion that falls in software and peripherals prices are inadequately captured in the national accounts. In fact, both are treated the same way as computers, with no quality correction to their price indexes.

[23] "Time Series Data for Input-Output Industries," BLS Publication 2018 (1979).

[24] See Ernst Berndt (1982) for conditions under which the existence of a well-defined price-performance index is guaranteed.

literature has recently been both summarized and substantially advanced by Robert Gordon (1985). I use two different price indexes in the calculations, the one from Gordon's preferred specification of computer quality and my extension of Chow's index forward to 1972. Both indexes use hedonic methods. I use my extension of Chow's price index because among the available indexes covering the period 1958–72, it shows the smallest fall, 19.08 percent per year in nominal terms. Gordon's index shows a rate decrease of 33.97 percent per year for that period. Given Gordon's observation that most of the important differences between the two price indexes occur before 1960, the two indexes reflect a realistic upper and lower bound on the price at which the demand curve for computers cuts the axis. Both indexes cover only the basic CPU and memory, not any peripherals, so they are consistent with my definition of the computer input.

A final reservation one might have about this application is that a great many things have been changing in the economic environment of the FSS other than the fall in the price of computing power. One class of changes in the economic environment can be captured by changes in income, or in the prices of other inputs or of other consumer goods. The flexible functional forms used here for preferences and for technology do not put restrictions on the elasticities of substitution between computers and other inputs in financial services, or on the income or price elasticity of demand for the output of the financial services sector. Price and income changes are well treated in index-number methods. If increased computerization is to be attributed to rising real wages in the financial services sector, or if the increase in the size of the sector is to be attributed to income-elastic demand for its output, such factors have been accounted for in the calculations.

Another class of changes in the economic environment is not "priced out." This includes demographic shifts which change demand, independent technical progress in the FSS, and so forth. Slightly stronger assumptions are needed here. The results of Caves et al. continue to hold if there are shifts in the first-order coefficients of the translog E^* between the two periods, but not shifts in the second-order coefficients. As a practical matter, this means that the model allows arbitrary shifts between the two periods in the desirability of any input or of any good, as long as these shifts do not change the demand elasticities. Thus independent technical progress in the FSS can be labor saving and computer using, but cannot change the elasticity of substitution between labor and computers. These are defensible assumptions, but, if they are false, then econometric methods could go beyond the index-number methods employed here to get a better answer.

B. *Biases to Cost Indexes*

This subsection reports calculations of the biases to the cost-of-living index and to a cost of (nonfinancial) production index from using the FSS price index from the national accounts. The downstream sector considered here is the private nonfinancial domestic economy. In the input-output system, it consists of the personal consumption expenditures component of final demand (*PCE*) and intermediate-input demand by all private business outside the FSS, which we call the private nonfinancial business economy (*PNFBE*).[25] These two sectors together we call the private domestic economy (*PDE*). The shares of the FSS in these sectors are shown in Table 3.[26]

Table 4 shows the results of calculating (4), the percentage bias to the conventional cost-of-living index caused by using the FSS implicit deflator as the price index for the sector.[27] The size of the bias is sensitive to

[25]In the 1972 *IO* table, the *PNFBE* is defined to include BEA sectors 1–69 and 71–76. In the 1958 table, it is 1–69, 71–76, and 81–83.

[26]It may be somewhat surprising that the share of the FSS in *PCE* is so much larger than in *PNFBE*. What is behind this is that the sales of the FSS to *PCE* are approximately equal to its entire sales as an intermediate input. Almost half of the intermediate-input sales of the FSS are to itself, so that after they are netted out the contribution to the *PNFBE* is considerably smaller.

[27]No important difference arises in calculating the real input cost index for the *PNFBE*. (See Caves et al.) However, some additional assumptions are needed to

TABLE 3—FSS SHARE IN DOWNSTREAM SECTORS' EXPENDITURES

	Downstream	
	1958	1972
PCE	.0407	.0529
PNFBE	.0253	.0213
Total *PDE*	.0320	.0358

TABLE 4—BIASES TO .5 $(C^{*0} \times C^{*1})$

Bias	*PCE*	*PNFBE*	*Entire PDE*
Gordon	.00587	.00257	.00408
Chow	.00384	.00172	.00269

the computer price index used. The economic importance of the estimated bias can be illuminated by an evaluation of the "outputs = inputs" approach to real price and output indexes. From 1958 to 1972, the implicit deflator for the FSS grew at 4.56 percent a year, on average. An alternative price index can be calculated by solving (4) for that $\hat{P}$ which gives a bias of zero. Economically, this is the price index implied by a perfectly competitive FSS with no internal technical progress. The annual rise in this price index was 3.98 percent (Chow index) or 3.67 percent (Gordon index). Thus either computer price index implies substantially faster real growth in the FSS than does the implicit deflator, even under the conservative assumption of no internal technical progress.

These figures are not out of line with other indexes of real output based on the "transactions" approach. Horst Brand and John Duke calculate a transactions-based real output index for commercial banks for the period beginning in 1967. Through 1972 it advances 1.5 percent per year faster than the deflator. Similarly, the transactions index of real output for the insurance sector used in the national accounts advances about 1.5 percent faster than the older, "liquidity" index.

C. *Spillover Calculations*

The size of the downstream welfare gains resulting from the fall in the price-performance ratio of computers can be estimated using (6). Results for this calculation are shown in Table 5 for each of the two price indexes and for three counterfactual shares of computers, s_Z^2. Given that the counterfactual is a 1972 world but with the 1958 relative price of computers, the base case (reported in the left column of the table) is $s_Z^2 = s_Z^1 = .91E^{-4}$. Many economic assumptions could have led to this share actually arising in 1972. For interpretational purposes, I prefer to use the assumptions of an expenditure elasticity of demand for computers of unity, and that computers are separable from everything else in E^*.[28] Then to get comparison cases, I maintain the separability assumption and assume that the expenditure elasticity of demand for computers is larger, taking on the values 2 and 3 in the second and third column of the table.[29]

The interpretation of these figures is as the amount by which consumers would be worse

TABLE 5—SPILLOVER CALCULATION: *PCE*

Case	(1)	(2)	(3)
s_Z^2	.000091	.000164	.000296
Spillover (Gordon) ($\$10^6$)	417	750	1351
Spillover (Chow) ($\$10^6$)	225	405	729
Expenditures ($\$10^6$)	68	68	68

calculate cost of living and of making Q_1 indexes for the two different customers of the FSS, final demand and intermediate demand. This calculation implicitly assigns the producer's surplus gained in the FSS, if any, to the two groups of customers in proportion to their sales. Further, the assertion that the presence of regulation leads to an underestimate of the size of the gains to each group of customers is obviously stronger than the result obtained in Section I. In particular, systematic cross subsidy between business and consumer customers of the FSS would invalidate that result.

[28]This would be true if the real income elasticity of demand for financial services is unity, the output elasticity of demand for computers in the FSS is also unity, computers are separable from other inputs in FSS production, and FSS outputs are separable from other goods in demand. These obviously unrealistic assumptions are used merely to provide an economic motivation for the nonbase cases.

[29]Income elasticities of demand for the FSS substantially greater than unity are reported by Kendrick (1982a).

off in 1972 if the real (rental) price-performance ratio for computers used in the FSS had not fallen from its 1958 level. For example, if we take the base case s_Z^2, consumers in 1972 would have been willing to pay \$225 million (Chow index) or \$417 million (Gordon) for this price decrease. Correspondingly higher estimates of this willingness to pay would be obtained if, for any reason, the share of computers in expenditures would have risen above its 1958 level even without the fall in computer prices. To put these figures in perspective, the amount spent by the FSS "for" consumers in renting computers in 1972 was \$68 million. Thus if we were to draw the derived demand curve for computers used in the FSS and put in the 1972 price, the consumer's surplus triangle would be five or more times the size of the $P \times Q$ box.

Similar figures would be obtained for the cost indexes of business customers of the FSS: the analog to Table 5 for them is essentially proportional to it.

I draw the following conclusion from these calculations: compared to expenditures on computers, the spillover to adopters of computers and their customers has been large. The present analysis has used a sector in which computers are particularly valuable, so that the finding might not generalize to the rest of the economy.[30] On the other hand, the price-performance ratio for computers has (conservatively) fallen another order of magnitude for computers since 1972. So in current (1986) terms, the downstream benefits of technical progress in mainframe computers since 1958 are conservatively estimated at 1.5 to 2 orders of magnitude larger than expenditures, at least in this high-value use.

Appendix: Proof of Result 1

Let λ be the Langrangian corresponding to $*1$, and let μ be the vector of Lagrangians for $*2$. Then the first-order conditions for the problem $**$ are

$$\text{(A1)}\quad Q_1:\ \lambda\frac{\partial U(Q_1,\Delta_1,Q_2)}{\partial q_j}=G_j-m_j+\mu_j \qquad j=1,\dots,R$$

$$Q_1:\ \lambda\frac{\partial U(Q_1,\Delta_1,\tilde{Q})}{\partial q_j}=\mu_j \qquad j=R+1,\dots,J$$

$$\text{(A2)}\quad Q_2:\ \lambda\frac{\partial U(Q_1,\Delta_1,Q_2)}{\partial q_k}=p_{2,k} \qquad \forall k$$

$$\text{(A3)}\quad Z:\ w_j=\mu_j\frac{\partial F_j(\delta_j,Z^j)}{\partial Z_n^j} \qquad \forall j$$

$$\text{(A4)}\quad \Delta_1:\ \lambda\frac{\partial U(Q_1,\Delta_1,Q_2)}{\partial \delta_j}=-\mu_j\frac{\partial F_j(\delta_j,Z^j)}{\partial \delta_j} \qquad \forall j$$

plus the constraints $(*1)$ and $(*2)$. Compare this system to the definitions of expenditure and cost functions:

$$\min P_1'Q_1+P_2'Q_2+\hat{\lambda}(u-U(Q_1,\Delta_1,Q_2))$$

$$\text{(A5)}\quad Q_1:\ \hat{\lambda}\frac{\partial U(Q_1,\Delta_1,Q_2)}{\partial q_j}=p_j \qquad \forall j$$

$$\text{(A6)}\quad Q_2:\ \hat{\lambda}\frac{\partial U(Q_1,\Delta_1,Q_2)}{\partial q_k}=p_{2,k} \qquad \forall k$$

$$\min W'Z^j+-\hat{\mu}_j\left(q_j-F_j\left(\delta_j,Z^j\right)\right)$$

$$\text{(A7)}\quad Z:\ w_j=\hat{\mu}_j\frac{\partial F_j(\delta_j,Z^j)}{\partial Z_n^j} \qquad \forall n$$

plus the constraints. From the envelope the-

[30] But see Manuel Tratjenberg's (1985) estimates of the external benefits of CAT scanners: Tratjenberg also provides an interesting discussion of the (difficult) welfare economics of technology adoption in the health sector.

orem we know that $\hat{\mu}_j$ is marginal cost and that

$$\frac{\partial E(u, P_2, P_1, \Delta_1)}{\partial \delta_j} = -\hat{\lambda}\frac{\partial U(Q_1, \Delta_1, Q_2)}{\partial \delta_j}$$

$$\frac{\partial C_j(q_j, \delta_j, W)}{\partial \delta_j} = \hat{\mu}_j \frac{\partial F_j(\delta_j, Z^j)}{\partial \delta_j},$$

so that (2) implies

$$\text{(A8)} \quad \hat{\lambda}\frac{\partial U(Q_1, \Delta_1, Q_2)}{\partial \delta_j} = -\hat{\mu}_j \frac{\partial F_j(\delta_j, Z^j)}{\partial \delta_j} \qquad \forall j.$$

At $mc_j = \mu_j$, for $j = 1, \ldots, R$, the sets of equations (A1)–(A4) and (A5)–(A8) are identical.

REFERENCES

Berndt, Ernst, "Quality Adjustment in Empirical Demand Analysis," mimeo., Sloan School of Management, MIT, 1982.

Brand, Horst and Duke, John, "Productivity in Commercial Banking: Computers Spur the Advance," *Monthly Labor Review*, December 1982, *105*, 19–27.

Caves, Douglas W., Christensen, Laurits R. and Diewert, W. Erwin, "The Economic Theory of Index Numbers and the Measurement of Input, Output and Productivity," *Econometrica*, November 1982, *50*, 1393–414.

Chow, Gregory C., "Technological Change and the Demand for Computers," *American Economic Review*, December 1967, *57*, 1117–30.

David, Paul A., *Technical Choice, Innovation, and Economic Growth*, London: Cambridge University Press, 1975.

Faucett, Jack (and Associates), *Regional Input-Output Tables for the United States*, Report to the U.S. Bureau of Labor Statistics, Washington, 1977.

______, *Development of Capital Stock Series By Industry*, Report to the U.S. Bureau of Labor Statistics, Washington, 1979.

Fogel, Robert W., *Railroads and American Economic Growth: Essays in Econometric History*, Baltimore: Johns Hopkins Press, 1964.

Fraumeni, Barbara and Jorgenson, Dale, "The Role of Capital in U.S. Economic Growth, 1948–1976," in G. von Furstenberg, ed., *Capital, Efficiency, and Growth*, Cambridge: Ballinger, 1980, ch. 3.

Gollop, Frank and Jorgenson, Dale, "U.S. Productivity Growth by Industry, 1947–1973," in J. Kendrick and B. Vaccara, eds., *New Development in Productivity Measurement and Analysis*, Chicago: University of Chicago Press, 1980, ch. 3.

______ **and** ______, "Sector Measures of Labor Cost for the United States, 1948–1978," Discussion Paper No. 963, Harvard Institute for Economic Research, 1983.

Gordon, Robert J., "The Postwar Evolution of Computer Prices," mimeo., Northwestern University, November 5, 1985.

Gorman, John A., "Alternative Measures of the Real Output and Productivity of Commercial Banks," in Victor Fuchs, ed., *Production and Productivity in the Service Industries*, NBER *Studies in Income and Wealth*, No. 34, New York: Columbia University Press, 1969, ch. 9.

Griliches, Zvi, "Issues in Assessing the Contribution of R&D to Productivity Growth," *Bell Journal of Economics*, Spring 1979, *10*, 92–116.

Jaffe, Adam, "The Effects of Market Demand, Technological Opportunity, and Research Spillovers on R&D Intensity and Productivity Growth," mimeo., Harvard University, 1984.

Hausman, Jerry A., "Exact Consumer's Surplus and Deadweight Loss," *American Economic Review*, September 1981, *71*, 662–76.

Kendrick, John, (1982a) *Interindustry Differences in Productivity Growth*, Washington: American Enterprise Institute, 1982.

______, (1982b) "Measurement of Output and Productivity in the Service Sector," presented at the ARA/Wharton Conference on the Future of the Service Economy, November 1982.

Mansfield, Edwin et al., *The Production and Application of New Industrial Technology*, New York: W. W. Norton, 1977.

Mark, Jerome A., "Measuring Productivity in Service Industries," *Monthly Labor Review*, June 1982, *105*, 3–8.

Marshall, Alfred, *Principles of Economics*, 8th ed., New York: MacMillan, 1948.

Phister, Montgomery, *Data Processing Technology and Economics*, Santa Monica: DEC Press, 1979.

Rosenberg, Nathan and Mowery, David, "Technical Change in the Commercial Aircraft Industry," in Nathan Rosenberg, ed., *Inside the Black Box: Technology and Economics*, Cambridge: Cambridge University Press, 1982, ch. 8.

Stigler, George, "Price and Non-Price Competition," *Journal of Political Economy*, January-February 1968, *76*, 149–54.

Tratjenberg, Manuel, "The Welfare Analysis of Product Innovations with Applications to CT Scanners," NBER Working Paper No. 1724, October 1985.

White, Lawrence, "Quality Variation when Prices are Regulated," *Bell Journal of Economics*, Autumn 1972, *3*, 425–36.

Willig, Robert D., "Sector Differentiated Capital Taxation with Imperfect Competition and Inter-Industry Flows," *Journal of Public Economics*, July 1983, *21*, 295–316.

Wyborg, Phillip S., McCarter, Pender M. and Erickson, William, *Information Processing in the U.S.: A Quantitative Summary*, Montvale: American Federation of Information Processing Societies, 1977.

U.S. Bureau of Labor Statistics, *Capital Stocks for Input-Output Industries*, Publication 2034, Washington, 1979.

______, *Time Series Data for Input-Output Industries*, Publication 2018, Washington, 1979.

Price Leadership and Welfare Losses in U.S. Manufacturing

By MICHA GISSER*

A model of price leadership is used to estimate the welfare losses due to monopoly in U.S. manufacturing. Given that the leaders behave independently, an hypothesis confirmed empirically, the deadweight loss is estimated at 0.114 percent of GNP based on a sample of 445 four-digit industries.

Thirty years ago, Arnold Harberger (1954) estimated the welfare loss due to monopoly in the United States. Since the appearance of Harberger's study, there has emerged a substantial literature on this subject. The articles by David Schwartzman (1960), David Kamerschen (1966), Frederick Bell (1968), Dean Worcester (1973), and John Siegfried and Thomas Tiemann (1974) constitute but a small sample out of the vast number of studies related to this topic. I believe that Sam Peltzman was correct when he suggested that "the reader has by this late day earned the right to demand strong justification for a new entrant" (1977, p. 229). My justification lies in the development of an innovative approach that applies to a model of price leadership in order to estimate the welfare loss.

It will be recalled that Harberger's analysis proceeds from an assumption of an infinitely elastic marginal cost that intersects with the industry-demand curve of unitary elasticity at a competitive price level. When the market price rises above the competitive level, consumers lose a portion of their consumers' surplus. Most of this loss is offset by the additional revenue that the producers obtain at the higher price, but a small portion of this loss is not offset by any gain to any group in the economy. The latter portion is the "deadweight" loss that has been estimated by Harberger and other researchers who basically applied the same method. But there is a distinct difference between the industry demand curve on the one hand, and the demand curve facing the individual firm or other maximizing agent (for example, a cartel) on the other. Most of the papers in this area have properly worked with the industry-demand curve in measuring the resource allocation costs of monopoly. But the relations between these curves and those facing the maximizing agents have been talked about explicitly only by Worcester (1975) and Harberger (1974, pp. 86–90). In both cases, moreover, the discussion is quite terse; it certainly has not been picked up by the subsequent published work in the area. It is this idea of the net-demand curve facing the individual firm, or the cartel of leaders, that is pursued in this article. A logical step in this direction would be to estimate the deadweight loss by utilizing a model in which the industry leaders confront a net-demand curve which is the horizontal difference between the demand for the industry and the supply of the small firms. This procedure can be justified on the basis of an empirical observation. I calculated the weighted average of firms per industry in an almost exhaustive sample of 445 4-digit industries taken from the 1977 *Census of Manufactures* at 1508, and the weighted average of the four-firm concentration ratio at 37.76. Value-added was used for weighting. These two statistics indicate that most manufacturing production is generated by industries with a concentration ratio lower than 50 percent, and these industries are composed of many small firms that are price takers.

Section I discusses the excess demand curve. In Section II, I estimate the dead-

*Professor of Economics, University of New Mexico, Albuquerque, NM 87131. I benefited substantially from comments by Peter Gregory and two referees of this journal. Any errors are mine.

weight loss under a price leadership model, first under the assumption that the four leading firms collude, and second under the assumption that they operate independently. To anticipate, the welfare loss is estimated separately for each of the 445 industries, and then it is aggregated and compared with the *GNP*. The order of magnitude of the welfare loss is higher than Harberger's estimate under the assumption that the leaders collude, but it is very close to Harberger's estimate under the assumption that the leaders act independently of each other. I also show that the deadweight loss decreases under the assumption that there are eight leaders instead of only four leaders, provided that they act independently of each other. Finally, Section III offers an indirect statistical test which supports the hypothesis that leaders act independently rather than collusively. I thus confirm the results of Harberger by utilizing an approach that more closely reflects the actual organization of the U.S. manufacturing sector than do the models used by others.[1]

I. Excess Demand

The central concept in this analysis is that of excess demand. The demand curve for an industry's product is one thing; that facing a collusion of leading producers is another; that facing an individual leading firm when there is no collusion is a third.

The basic arithmetic of excess demand is shown below. Excess demand (E) is defined as the difference between total demand (D_T) and other supply (S_0). Thus

$$(1)\quad E \equiv D_T - S_0$$

$$(2)\quad \frac{P}{E}\frac{\partial E}{\partial P} = \frac{D_T}{E}\left[\frac{P}{D_T}\frac{\partial D_T}{\partial P}\right] - \frac{S_0}{E}\left[\frac{P}{S_0}\frac{\partial S_0}{\partial P}\right]$$

$$(3)\quad \eta_E = (D_T/E)\eta_T - (S_0/E)\varepsilon_0.$$

In an industry with a homogeneous product, each producer "sees," as the demand curve facing him, an excess curve whose elasticity is governed by the formula (3).

Thus if the elasticity of demand for the *industry's* product is -1, and if each firm's marginal cost schedule has an elasticity of $+1$, then an individual firm accounting for 10 out of an industry's output of 100 will perceive its demand curve as having an elasticity η_E of $-19[=(100/10)(-1)-(90/10)(+1)]$. A cartel consisting of four such firms will see the elasticity of excess demand facing it as $-4[=(100/40)(-1)-(60/40)(+1)]$.

We know from the theory of the firm that a profit-maximizing entity will try to equate marginal revenue with marginal cost. Thus, an industry with four leading firms, each accounting for 10 percent of output and each setting price in accordance with some oligopolistic mode of behavior, will tend to a situation in which, for the leading firms, marginal cost is about 94 percent of market price. If the same four firms collude, they will end up operating where marginal cost is three-fourths of market price.[2]

The efficiency cost (deadweight loss) of the monopoly distortion can in each of these two cases be measured along the excess demand curve facing the group of four leading firms (as even in the case where they do not collude, the leading firms end up each doing the same thing). In the collusion case the welfare cost is a triangle whose height is 1/4 of demand price, in the case of no collusion the welfare cost triangle has a height of only 1/16 of demand price.

[1]This study does not address the issue of dynamic aspects of U.S. manufacturing studied by John Kendrick and Elliot Grossman (1980), Peltzman, Steven Lustgarten (1979), and myself (1982, 1984).

[2]Let η_E denote the elasticity of the net demand curve. Then, if the 4 leaders collude, for them $MR = P(1+1/\eta_E) = P\cdot 0.75$. If the 4 leaders act independently, some oligopolistic mode of behavior must be assumed. I opted for the Cournot mode simply because it is mathematically convenient. Then the elasticity perceived by each of the 4 leaders is $4\cdot\eta_E$, and

$$MR = P[1+1/(4\cdot\eta_E)]$$
$$= P[1+1/(-16)] = P\cdot 0.94.$$

TABLE 1—SIMULATED PRICE DISTORTIONS AND DEADWEIGHT LOSS IN PERCENTS UNDER COLLUSION AND INDEPENDENT (COURNOT) MODE OF BEHAVIOR

Four-Firm Concentration	η Total-Demand Elasticity	ε: Elasticity of Supply of Price Takers and Aggregate MC Curve of Price Leaders							
		1		2		10		100	
		Collusion	Cournot	Collusion	Cournot	Collusion	Cournot	Collusion	Cournot
20	−1	10.132	2.711	6.756	1.858	1.851	0.527	0.203	0.058
		[0.094]	[0.007]	[0.078]	[0.006]	[0.027]	[0.002]	[0.003]	[0.000]
	−2	6.727	1.759	5.063	1.355	1.696	0.477	0.200	0.058
		[0.043]	[0.003]	[0.046]	[0.003]	[0.023]	[0.002]	[0.003]	[0.000]
40	−1	20.871	5.966	14.090	4.233	3.909	1.266	0.428	0.142
		[0.756]	[0.058]	[0.601]	[0.053]	[0.188]	[0.020]	[0.021]	[0.002]
	−2	13.844	3.745	10.531	2.985	3.581	1.128	0.424	0.140
		[0.351]	[0.025]	[0.363]	[0.029]	[0.165]	[0.017]	[0.021]	[0.002]
60	−1	33.333	10.070	22.908	7.492	6.506	2.413	0.718	0.278
		[2.847]	[0.225]	[2.143]	[0.211]	[0.604]	[0.081]	[0.065]	[0.010]
	−2	21.922	6.055	17.026	5.043	5.957	2.102	0.710	0.273
		[1.312]	[0.091]	[1.340]	[0.110]	[0.551]	[0.067]	[0.064]	[0.009]
80	−1	50.000	15.693	35.604	12.613	10.601	4.619	1.186	0.555
		[9.153]	[0.661]	[6.473]	[0.659]	[1.550]	[0.261]	[0.153]	[0.030]
	−2	32.195	8.856	26.102	7.866	9.693	3.874	1.175	0.544
		[3.945]	[0.245]	[4.130]	[0.318]	[1.506]	[0.217]	[0.153]	[0.029]

Note: Distortions and losses (shown in brackets below) are in percents of price and sales (of price leaders and price takers combined), respectively, They are calculated by applying numeric methods.

Since the welfare costs vary roughly with the square of the relevant distortion, assuming four large firms, one can expect the cost under cartel behavior to exceed those with no collusion by a factor equal to approximately $(16)^2/(4)^2$ or 16. In general, this factor will be equal to the square of the number of leaders.[3]

Table 1 presents simulated distortions and deadweight losses in percents of price and total value of industry output, respectively, under collusion and under a independent (Cournot) mode of behavior. A numeric method that is described in the next section is applied in calculating the price and quantity distortions. The loss defined as $-0.5 \cdot \Delta P \cdot \Delta Q$ of the price leaders, is divided by total value of the industry output of both leaders and price takers in order to convert it from a dollar amount to a percentage. The results in Table 1 are thus not exactly equal to results that would be obtained by utilizing equation (3). Had equation (3) been used, a point would be located on a line which is tangent to the excess demand curve at the competitive price. The numeric method selects an equilibrium point on the excess demand curve as defined by equation (1). The simulation results in Table 1 show that selecting for marginal costs and demand curves unitary elasticities, as compared to higher elasticities, biases the results in the direction of larger welfare losses. The losses do not change appreciably when the marginal cost elasticity rises from 1 to 2. They are cut by roughly one-half when the marginal cost elasticity rises to a magnitude of 10. The losses are very sensitive to changes in the total demand elasticity. Increasing the total

[3]The deadweight loss may be measured by the formula

$$\text{Loss} = \frac{1}{2}\left(\frac{\Delta P}{P}\right)^2 \cdot R \cdot \left[\frac{1}{1/\eta_E + 1/\varepsilon}\right]$$

where R is the value of output produced by the large firms, and ε is the elasticity of the marginal cost curves. Let M denote the number of large (equal-sized) firms. If there are 4 large firms and assuming a Cournot mode of behavior, then $M = 4$. Since (Lerner's ratio) $\Delta P/P =$ 1/elasticity, we obtain a factor of

$$(1/\eta_E)^2/[1/(M \cdot \eta_E)]^2 = M^2.$$

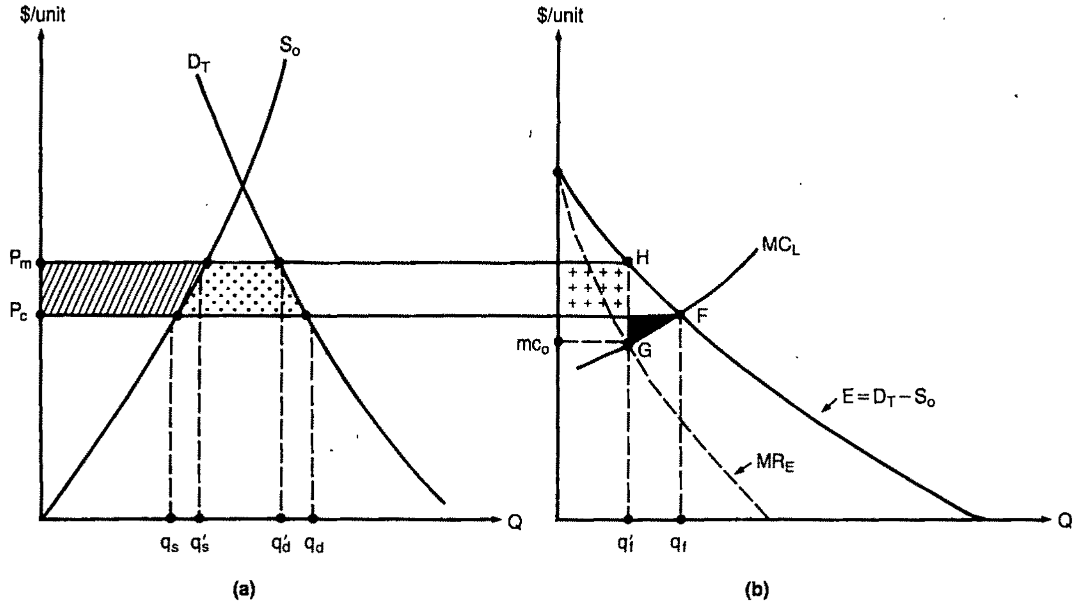


FIGURE 1. A PRICE LEADERSHIP MODEL

demand elasticity (absolute value) from -1 to -2 results in roughly 50 percent loss reduction. As expected, the loss rises rapidly as the concentration ratio increases. For example, if a Cournot mode of behavior coupled to unitary elasticities is adopted, the loss expressed as percentage of the total value of industry output rises from 0.007 to 0.661 when the four-firm concentration ratio rises from 20 to 80 percent.

II. Deadweight Loss and Price Leadership

The model of price leadership, or dominant firm as some prefer to call it, is summarized in Figure 1. In order to obtain estimates of the deadweight losses for each of the 445 industries included in our study, the quantitative aspects of the model that were discussed partially in the previous section are further developed in this section.

The aggregate demand D_T is

$$D_T = HP^{\eta_T}, \tag{4}$$

where D_T, H, P, and η_T stand for total output demanded, demand shifter, price, and price elasticity of the total demand curve.

The supply of price takers S_0 is

$$S_0 = G_s P^{\varepsilon_0}, \tag{5}$$

where S_0, G_s, P, and ε_0 represent the quantity supplied by price takers, supply shifter, price, and price elasticity of the supply of price takers.

The aggregate marginal cost of the leaders (or dominant firm) S_L is

$$S_L = G_f (MC_L)^{\varepsilon_L}, \tag{6}$$

where S_L, G_f, MC_L, and ε_L represent the quantity produced by the leaders, aggregate marginal cost (of the leaders) shifter, marginal cost of the leaders, and the elasticity of the aggregate marginal cost of the leaders.

The net-demand curve confronting the leaders (or dominant firm) is derived as

$$E = D_T - S_0. \tag{1'}$$

Note that although the elasticities along curves D_T and S_0 in Figure 1(a) are respectively constant, the elasticity of curve E in Figure 1(b) is not constant.

The marginal revenue curve associated with the net-demand curve confronting the leaders is denoted by MR_E. Assuming collusion among the leaders, point G in Figure 1(b), where MR_E cuts MC_L, would determine the optimal level of output for the leaders. Accordingly, they will agree on a price of P_m and produce a quantity q_f'. A price P_c and a quantity q_f would prevail in the market if the leaders were to behave competitively. In general, quantities with super prime are associated with the price P_m. Half of the price distortion (GH) multiplied by $q_f' - q_f$ is the welfare loss. It is the area of the triangle HGF in Figure 1(b).[4]

If the leaders were to act independently of each other, the quantity produced by them in the aggregate would be determined at a point between q_f' and q_f (not shown in Figure 1), where a more elastic marginal revenue curve associated with a more elastic demand curve intersects with the aggregate marginal cost curve of the leaders denoted by MC_L. The procedure of calculating the deadweight loss to society would then be the same as in the case of collusion.

In order to generate empirical estimates of deadweight losses due to collusion or any other mode of oligopolistic behavior, numerical values have to be assigned to parameters in equations (4)–(6). I assign a value ranging from -1 to -10 to the price elasticity of the total demand curve (η_T). Perfectly elastic marginal cost curves were assumed by Harberger and others. Since horizontal cost curves are cumbersome in price leadership models, positively sloping curves with constant elasticity ranging in value from 1 to 10 are assumed. For numerical convenience I assume that the competitive price (P_c) is \$1, and the quantity marketed at that price is 100 units. Thus the aggregate demand coefficient H is set at 100, and if the concentration ratio is denoted by C, then the coefficient G_s of the supply of the price takers is set at $(100-C)$ while the coefficient G_f of the aggregate marginal cost of the price leaders is set at C. I also assume that $\varepsilon_0 = \varepsilon_L = \varepsilon$. Assuming a Cournot mode of behavior among the M leaders, the noncompetitive price can be calculated by solving numerically for the following pair of simultaneous equations:[5]

$$(7)\quad E = 100P^{\eta_T - 1} - (100-C)P^{\varepsilon}$$

$$(8)\quad (E/C)^{1/\varepsilon} = (E/M)\left[1/\left(100\cdot\eta_T\cdot P^{\eta_T-1} - (100-C)\cdot\varepsilon\cdot P^{\varepsilon-1}\right)\right] + P.$$

The parameter M is set equal to 4 if it is assumed that the four leaders follow a Cournot mode of behavior, and it is set at 1 if it is assumed that the leaders collude. Once the noncompetitive price is calculated, the noncompetitive quantity produced by the four large firms is calculated from equation (7). This new quantity minus the initial rate of production of the large firms is equal to the segment $q_f' - q_f$ in Figure 1(b). The mar-

[4]The welfare loss can also be derived from a consideration of the following elements in Figure 1: Loss to consumers (area covered by diagonal stripes plus dotted area in panel a): $(P_m - P_c)[(q_d + q_d')/2]$; Gain to price takers (area covered by diagonal stripes in panel a): $(P_m - P_c)[(q_s + q_s')/2]$; Gain to price leaders (rectangle covered by crosses in panel b): $(P_m - P_c)q_f'$; Loss to price leaders (shaded triangle in panel b): $(P_c - mc_0)(q_f - q_f')/2$. Deadweight loss to society is the loss to consumers plus the loss to price leaders, minus the gain to all producers.

[5]If we calibrate the price-leadership model such that at a competitive price of \$1 the quantity cleared in the market is 100 units, and additionally if we substitute equations (4) and (5) into equation (1) we obtain equation (7), where $100-C$ replaces G_s, and C is the concentration ratio. Assuming M leaders, then $E = \Sigma_{i=1}^{M} Q_i$, and $dE/dQ_i = 1$. Total revenue of the ith leader under a Cournot mode of behavior is $R_i = Q_i P$. Differentiating with respect to Q_i we obtain

$$MR_i = dR_i/dQ_i = (E/M)(dP/dE) + P,$$

where E/M is equal to Q_i under the assumption that the leaders have equal sizes. The marginal cost curve of a single leader obtained by dividing equation (6) by M and, recalling that $G_f = C$ and $\varepsilon_L = \varepsilon_0 = \varepsilon$, is $S_L/M = (C/M)(MC_L)^{\varepsilon}$, which may be expressed as $MC_L = (S_L/C)^{1/\varepsilon}$. Since in equilibrium $MC_L = MR_i$ ($MC_L \equiv MC_i$), and $S_L = E$, we obtain equilibrium condition (8). This equation may be solved numerically after substituting $100P^{\eta_T} - (100-C)P^{\varepsilon}$ for E. The expression in the square brackets is dP/dE.

TABLE 2—DEADWEIGHT LOSS IN DOLLARS AND AS A PERCENTAGE OF *GNP*: 1977

	Collusion					Cournot				
	$\eta=-1$ Deadweight Loss			$\varepsilon=1$ DeadWeight Loss			$\eta=-1$ Deadweight Loss			$\varepsilon=1$ Deadweight Loss	
ε	Million Dollars	Percent of *GNP*	η	Million Dollars	Percent of *GNP*	ε	Million Dollars	Percent of *GNP*	η	Million Dollars	Percent of *GNP*
1	34,444.5	1.823	−1	34,444.5	1.823	1	2,156.8	0.114	−1	2,156.8	0.114
2	24,216.9	1.282	−2	13,445.1	0.712	2	2,232.9	0.118	−2	790.7	0.042
4	13,656.6	0.723	−4	4,152.2	0.220	4	1,771.9	0.094	−4	249.5	0.013
10	5,548.9	0.294	−10	762.3	0.040	10	920.0	0.049	−10	46.8	0.002

Note: In 1977 total shipments of manufacturing industries amounted to $1,287.118 billion. My sample of 445 4-digit industries accounted for $1,267.983 billion. Hence, the dollar figures of losses were blown up by a factor of 1.015 before calculating the losses as percentages of *GNP*. In 1977, *GNP* was $1,918.0 billion.

ginal cost of the large firms is obtained by substituting the calculated noncompetitive price and quantity (of the large firms) into equation (8). The calculated price minus the calculated marginal cost of the large firms is depicted by the segment $P_m - mc_0$ in Figure 1. The deadweight loss is one-half the product of the two calculated segments. The deadweight loss as percentage of the value of total output is obtained by dividing it by the expression $H\hat{P}^{\eta_T+1}$ where $\hat{P}$ is the calculated price, and H is set equal to 100. Price $\hat{P}$ is depicted by P_m in Figure 1 if collusion is assumed.

A sample of 445 4-digit SIC manufacturing industries taken from the 1977 *Census of Manufactures* was used in order to estimate the deadweight loss. The description of the data is set forth in the Data Appendix. Given a pair of η_T and ε, I calculated for each of the 445 industries its percentage deadweight loss based on its four-firm concentration ratio, and then applied this percentage to the industry's total value of shipments in order to obtain its dollar value of deadweight loss.[6]

Finally, I aggregated the dollar values of the deadweight losses of the 445 industries, and divided it by the 1977 *GNP* in order to obtain the deadweight loss as percentage of *GNP*. The results are contained in Table 2.

Under the assumption that the leaders collude and $\eta=-1$, the deadweight loss is 1.823 percent of *GNP* for $\varepsilon=1$, and only 0.294 percent for $\varepsilon=10$. Under the assumption that the leaders act independently, that is, in my model they follow a Cournot mode of behavior, and $\eta=-1$, the deadweight loss is 0.114 percent of *GNP* for $\varepsilon=1$, and only 0.049 percent for $\varepsilon=10$. If demand elasticity is allowed to increase beyond -1, the estimate of the loss decreases very rapidly. It seems to me that while a unitary demand elasticity would be acceptable as a

[6]As an illustration, consider industry 2034 with a concentration ratio of 37 percent and value of shipments of $1142.4 million. Assume $\eta_T=-1$, $\varepsilon=1$; recall that initially competition prevails and $P_c=\$1$, and total output is set at 100 units. Applying the numeric methods that were discussed in the text the vertical difference between the oligopolistic price and the marginal cost of the independent leaders is calculated at 0.055 and the change in the rate of production of the leaders is calculated at 1.653. The loss is then $-(1/2)\cdot 0.055\cdot 1.653=-0.04546$. Since $\eta_T=-1$ the loss figure is not deflated by $H\hat{P}^{\eta_T+1}$. The dollar value of the loss is $(-0.04546\cdot 1142.4)/100=-\$0.518$ million. Consider now the "crude" method of estimating the welfare loss associated with industry 2034. Utilizing equation (3), the elasticity of the net-demand curve E is $\eta_E=(100/37)(-1)-(63/37)(1)=-4.4054$. The elasticity perceived by an independent leader ($M=4$) in a Cournot environment is $-4.4054\cdot 4=-17.622$. Thus, the loss as a percentage of sales is

$$(1/2)\left(\frac{1}{17.622}\right)^2\cdot 37\cdot\left[\frac{1}{1/(-4.4054)-1/1}\right]=0.048.$$

Note that the crude 0.048 is about 5 percent higher than 0.04546.

TABLE 3—CONTRIBUTION TO THE DEADWEIGHT LOSS BY THE FIVE LEADING INDUSTRIES

SIC Code	Industry Description	4-Firm Concentration Ratio	Value of Shipments $\$\times10^6$	Deadweight Loss $\$\times10^6$
3711	Motor Vehicles and Car Bodies	93	76,517.8	993.873
3714	Motor Vehicle Parts and Accessories	62	35,750.8	90.327
3861	Photographic Equipment and Supplies	72	9,946.9	43.411
2824	Organic Fibers, Noncellulosic	78	6,379.7	38.041
3011	Tires and Inner Tubes	70	8,971.0	35.212

Note: $\Sigma = 1200.864$.

rough representation of real life, supply elasticity could vary over a wide range extending from unit to infinity. Thus, the estimate of 1.823 percent under collusion and 0.114 percent under a Cournot mode of behavior are adopted as the two upper limits of the estimates of the deadweight losses, respectively. The interesting result is that if the leaders do not collude, the upper limit of my estimate proves to be very close to Harberger's one-tenth of 1 percent of national income calculated for the late 1920's.

As shown in Table 3, five industries account for $1,200.864 million of the loss. This is 55 percent of the total monopoly loss of $2,156.8 million. More important, motor vehicles (SIC 3711) and motor vehicle parts (SIC 3714) together account for $1,084.2 million, which is half of the total loss. This result confirms what Siegfried and Tiemann discovered earlier.

My next step was to assume that in industries with a concentration ratio lower than 50 percent, the leaders follow a Cournot mode of behavior; in industries with a concentration ratio equal to or exceeding 50 percent, the largest firm dominates half of the share of the four leaders. It acts as the dominant firm by setting a monopolistic price based on the net demand curve confronting it; the rest of the firms equate their marginal cost curves with this price. The result is a deadweight loss of 0.124 percent of *GNP*.

The Census provides the eight-firm concentration ratio. What if we assume that there are eight leaders who follow a Cournot mode of behavior? The answer is that, assuming unitary elasticity, for a sample of 440 industries, the deadweight loss amounted to 0.054 of 1 percent of *GNP*. The Data Appendix explains why 5 industries were omitted from the first sample of 445 industries.

It should be noted that in the earlier analysis, it was necessary to ignore the cross-price effects; accordingly this is a strictly partial-equilibrium analysis. That is, I assume that the demand curve for each industry is independent of the prices of goods in other industries. Ignoring the cross-price effects among the manufacturing industries biases upwards the partial-equilibrium estimate. It is as if each industry were being subjected to a tax imposed and collected by its leading firms. When we look at the actual data of a real economy, we have many different industries, in each of which the leading firms are thought of as imposing and collecting a tax—a relatively large tax when they collude, a relatively smaller tax when they do not. In this case, there are interactions resulting from the simultaneous consideration of different distortions in the different industries. When all the distortions are positive (as I assume) and where the relations among most goods are characterized by sub-

stitution (i.e., no complementarity), the interaction necessarily results in the reduction of the overall welfare cost.[7] The size of this reduction will vary from case to case. It depends not only on the average size of the distortions being imposed, but also on their variance and on the fraction of the economy covered by the "monopolized sector" (i.e., by that group of industries for which monopoly distortions are being measured and their welfare cost assessed).

Ignoring the relationship between the aggregate sector of manufacturing industries on one hand, and the rest of the economy on the other hand, might bias the welfare loss estimate slightly. The reason for this is that shifting of inputs from the firms with monopoly power to the rest of the economy will result in a small reduction in the prices of these inputs. An exercise in general equilibrium shows that such industrywide effects biases the partial-equilibrium estimate upward by a very small amount.[8]

In the following section, I perform a statistical test that lends support to the hypothesis that the leaders are not engaged in collusive agreements, or if they do they engage in "secret" price cutting, such that their behavior is roughly approximated by a Cournot-Nash model. In that case, my results will serve to confirm Harberger's findings.

III. Statistical Inference Against Collusion

Which of the extreme values of the estimated loss is a closer reflection of reality

[7]If following Harberger (1974, ch. 2), one thinks of the distortions being imposed sequentially in activities 1 through k ($k=445$), the first distortion generates a triangle of welfare cost. The second generates its own triangle of cost plus a rectangle of welfare gain due to the necessarily rightward shift of the demand for the first commodity induced by the imposition of the second distortion. The third distortion produces its own cost triangle, but now there are two rectangles of gain, owing to the rightward shifts of demand in the two already distorted activities. When the kth distortion is introduced, we get the final triangle of cost, this one accompanied by $k-1$ rectangles of gain. It is possible under certain circumstances that the sequence of rectangles of gain thus generated will cancel (or even outweigh) the losses represented by the triangles, but this is not possible in the exercises performed in this paper where it is assumed (a) that one starts from an undistorted initial equilibrium, and (b) that large segments of the economy (service activities, etc.) are uncovered by the set of positive distortions imposed.

[8]First, consider the partial-equilibrium analysis. Initially the four potential leaders produced 40 percent of output X leaving 60 percent to the price takers. The competitive price is \$1. After they collude, the leaders raise the price to \$1.0455. The partial-equilibrium welfare loss due to collusive leadership is as follows: Loss to consumers, -4.359; Gain to price takers, 2.733; Gain to leaders, 1.473; Loss to leaders, -0.602; Deadweight loss, -0.754. Since manufacturing output amounts to 25 percent of GNP, the figure 0.754 translates into $0.754/4=0.19$ percent of GNP. Second, consider a general equilibrium setting in which the utility is identical for all consumers and is homogeneous of degree one in two goods, X and Y. In particular, let the utility function be $U=Y^{0.75}X^{0.25}$, where X is the output produced by the manufacturing sector, and Y is the good produced competitively by the rest of the economy. Such a utility function agrees with the fact that manufacturing output amounts to 25 percent of GNP. The homogeneity of the utility function guarantees that the welfare loss as indicated by the percentage change in the level of utility is equal to the increase in national income that would be necessary, at the monopoly prices, for consumers to enjoy the competitive utility level. Finally, the demand curves associated with this utility function are of unitary elasticity. Let X_t and X_d denote the output of the price takers and the leaders in the manufacturing sector. Let A denote the amount of some variable input used in the production of both Y and X. The three production functions are

$$Y=1.22474\cdot A^{0.5};\ X_t=0.54772\cdot A^{0.5};\ X_d=0.44721\cdot A^{0.5}.$$

These production functions give rise to marginal cost curves that are of unitary elasticities. Initially, the leaders produce 40 percent of X, but they sell it in a competitive market. The values of P_a, P_y, P_x, Y, X_t, X_d, MC (of X_d), and U are 0.01, 1, 1, 75, 15, 10, 1, and 56.988, respectively. They then collude and raise the price of X until their marginal revenue is equal to the marginal cost that they perceive, namely the ratio of the price of A over the marginal physical product of A, in the production of X_d. The dominant firm solution is obtained by applying numeric methods. The results for P_a, P_y, P_x, Y, Y_t, X_d, MC (of X_d), and U are 0.00982, 0.991, 1.035, 75.685, 15.811, 8.341, 0.819, and 56.885, respectively; the utility falls by 0.18 percent. Note that the deadweight loss in a model of general equilibrium of 0.18 percent of GNP is slightly less than the partial-equilibrium deadweight loss of $0.754/4=0.19$ percent of GNP. Thus, we may feel confident that our estimate of the welfare loss due to monopoly power in the manufacturing sector is upward biased on account of industrywide effects.

depends on which mode of behavior is thought to be more relevant, collusive or independent. I believe that a strong case can be made in favor of a Nash solution in which each oligopolist behaves optimally given the decisions of his other rivals.[9] Such noncollusive solutions are appropriately identified not simply as Cournot but rather as leading to a Cournot-Nash equilibrium. Nevertheless, a measure of caution is justified. As George Stigler has pointed out,

> the literature of collusive agreements, ranging from the pools of the 1880s to the electrical conspiracies of recent times, is replete with instances of the collapse of conspiracies because of "secret" price cutting. This literature is biased: conspiracies that are successful in avoiding an amount of price cutting which leads to collapse of the agreement are less likely to be reported or detected. [1964, p. 46]

The null hypothesis is that the leaders collude. It would be ideal if it were possible to test the null hypothesis by relating empirically the deadweight loss to concentration. But the deadweight loss is not observable. Accordingly, I calculate for each industry the supernormal profit which is closely associated with the deadweight loss, and test its correlation with the concentration ratio. Using data from the 1977 *Census of Manufactures*, the cost is defined as the sum of total labor cost, the average of assets and inventories at the beginning and end of the year multiplied by the rate of interest, depreciation charges, and rental payments. Profit is then defined as value-added minus cost. To standardize the profit variable I divide it by value-added. The rate of interest chosen is 5 percent. A detailed description of the data which contain 445 industries is provided in the Data Appendix.

The result of regressing the standardized profit (Π) on the four-firm concentration ratio (C) is summarized by the following

[9]For an excellent exposition, see Jack Hirshleifer (1974, ch. 10).

TABLE 4—SIMULATED PROFIT DEFLATED BY THE VALUE OF PRODUCTION UNDER COLLUSION OF PRICE LEADERS[a]

4-Firm Concentration Ratio	ε: Elasticity of Supply of Price Takers and Aggregate Marginal Cost Curve of Price Leaders		
	1	10	100
10	0.2393	0.0765	0.0089
20	0.9269	0.2891	0.0340
30	2.0435	0.6275	0.0733
40	3.6038	1.0986	0.1283
50	5.6624	1.7265	0.2016
60	8.3333	2.5645	0.2998
70	11.8410	3.7214	0.4364
80	16.6667	5.4549	0.6438
90	24.1497	8.6119	1.0306

[a]Demand elasticity equals -1.

equation:

$$\Pi = \underset{(22.36)}{0.332} + \underset{(3.02)}{0.001C} + e, \tag{9}$$

where the t-ratios are shown in parentheses, the F-ratio is 9.05, and the R-square is 0.02. The estimate of the intercept is very significant. It tells us that, *ceteris paribus*, manufacturing industries could expect that profit as defined above would amount to 33 percent of value-added. The slope of the concentration ratio is very small but significant. It tells us that profit would rise by one-tenth of 1 percent per one-percentage-point rise in the concentration ratio. As an example, if the concentration ratio of an industry were to increase from 40 to 70 percent, profit would increase by 3 percent of value-added.

The null hypothesis is that the leaders collude. For an industry with a demand elasticity of -1 and supply and aggregate marginal cost of various elasticities, I simulated the profits deflated by the value of production for nine levels of concentration. The results of the simulation are shown in Table 4. Regressing the simulated deflated profits (col. 2, $\varepsilon = 1$) on the concentration ratios (col. 1) yielded the following equation:

$$\Pi = \underset{(-2.83)}{-0.058} + \underset{(7.68)}{0.003C}. \tag{10}$$

Under the assumption of unitary supply elasticity and collusion, standardized profit would increase by 9 percent if the concentration ratio of such an industry were to increase from 40 to 70 percent. This is three times as great as that estimated by equation (9) which reflects the real world.

If the supply elasticity is increased to 10, then regressing the simulated deflated profits (col. 3, $\varepsilon = 10$) on the concentration ratios yields the following:

$$\Pi = -0.021 + 0.001C. \tag{11}$$
$$(-2.48) \quad (6.38)$$

The slope of the concentration ratio in equation (11) is the same as the slope in equation (9). The significance of this result is that if we assume that the supply elasticity is equal to 10, we cannot reject the null hypothesis that the leaders collude. However, a glance at Table 1 tells us that increasing the supply elasticity from unity to 10 would result in cutting the deadweight loss by three-fourths: For a concentration ratio of 40 percent, we have $0.188/0.756 = 0.249$. That is, as shown in Table 2, instead of an estimate of 0.114 of 1 percent of the value of output under the assumption of supply curves of unitary elasticity and a Cournot mode of behavior, the deadweight loss under the assumption of supply curves with an elasticity of 10 and a collusive behavior would amount to 0.249 of 1 percent of *GNP*. This seems to be the upper limit of the deadweight loss to society due to monopoly power in manufacturing industries.

IV. Summary

Harberger (1954) and researchers who followed in his footsteps estimated the deadweight loss to society from monopoly power in manufacturing industries by applying a model in which the marginal cost curve is infinitely elastic, the aggregate demand curve is of unitary elasticity and the loss is the area of the triangle under the demand curve formed by fitting the difference between the observed price and calculated marginal cost. My study was motivated by the simple empirical finding that in most manufacturing industries output is produced by many small firms and a few giants that coexist side by side. For an almost exhaustive sample of 445 4-digit SIC industries taken from the 1977 *Census of Manufactures*, the weighted average of the number of companies was calculated at 1508, and the weighted average of the four-firm concentration ratio at 37.76. These two statistics lend strong support to an observation made by Worcester (1975) and Harberger (1974) that oligopoly theory implies that welfare losses should be based on more elastic demand curves facing individual firms, rather than on industry demand curves. I extended this idea to a price leadership model in which the leaders confront a net demand curve derived by subtracting the supply curve of the small firms from the total industry demand curve. For purposes of comparison I selected industry demand curves of unitary elasticity. Since infinitely elastic marginal cost curves are not appropriate for a price leadership model, I opted to assume finite elasticities. In comparison with earlier research, this assumption biases the results towards an overestimation of the welfare loss since the less elastic the supply curve of the price takers and the marginal cost curves of the leaders, the larger the deadweight loss. Although the ideas advanced by Stigler coupled to a Cournot-Nash process of reaching equilibrium convince me that the leaders behave like independent oligopolists, I estimated and contrasted deadweight loss for the sample of 445 industries under the assumption of collusion and under the assumption of a Cournot mode of behavior. The collusion estimate serves as a null hypothesis to be tested.

Under the assumption that the leaders collude, the deadweight loss was estimated at 1.823 percent of *GNP*, and under the assumption of a Cournot mode, it was estimated at 0.114 percent of *GNP*. These two estimates are derived under the assumption that the supply elasticity is equal to unity. By regressing profits deflated by value-added on concentration ratios in a sample of 445 industries I could reject the collusion hypothesis provided that the supply elasticity is equal to one. If supply elasticity is on the order of magnitude of 10 or more, the collu-

sion assumption cannot be rejected, but then the order of magnitude of the deadweight loss to society is 0.294 percent of *GNP*, enough to treat every family to two steak dinners and also close to Harberger's estimate. Assuming that industries with a concentration ratio in excess of 50 percent behave according to a single dominant-firm model, where the single firm has one-half of the four-leader share, and the industries with a concentration ratio smaller than 50 percent follow a Cournot mode, resulted in a loss estimate of 0.124 percent of *GNP*. Finally, utilizing a sample of 440 industries, I estimated the deadweight loss under the assumption that the eight largest firms in each industry follow a Cournot mode at 0.054 percent of *GNP*. This is the lowest estimate in the set.

My conclusion is that estimates of the deadweight loss to society derived from an industry model characterized by price leadership lend strong support to Harberger's findings of thirty years ago.

DATA APPENDIX

All data are taken from the U.S. Bureau of the Census, *Census of Manufactures*, 1977, Vol. 1.

Number of firms (companies), concentration ratios (C-4, and C-8) and total shipments are obtained from *Concentration Ratios in Manufacturing*, Table 7.

Value-added by manufacture, rather than value of shipments are given for industries 2011, 2013, 2271, 3312, 3331, 3332, 3573, and 3585, because their values of shipment contain a substantial and unmeasurable amount of duplication.

Labor costs figures are obtained from *General Summary*, Table 2.

Value-added and inventories at the beginning and end of year, respectively, are taken from *General Summary*, Industry Statistics.

Assets at the beginning and end of year, depreciation charges and rental payments are obtained from *Gross Book Value of Depreciable Assets, Capital Expenditures, Retirements, Depreciation and Rental Payments*, Table 2.

The first sample contains 445 4-digit SIC industries out of an exhaustive list of 450 industries. This sample is used in analysis focusing on the four-firm concentration ratios. Industries 2111, 2823, 3572, 3647, and 3661 were not included because the Census does not provide the four-firm concentration ratio for these industries. Industry 3332 has a concentration ratio of 100 percent. Since I assume an aggregate demand curve of unitary elasticity and a model of price leadership, I had either to drop that industry or slightly change its concentration ratio, I opted to lower its concentration ratio to 95 percent.

The second sample contains 440 4-digit SIC industries drawn from the first sample. The second sample is used in analysis focusing on the eight-firm concentration ratio. It is obtained by deleting industries 2661, 3031, 3263, and 3633 for which the Census does not provide the eight-firm concentration ratio. In industries 2296, 2895, 3331, and 3333, the eight-firm concentration ratio was changed from 100 percent to 95 percent.

The omission of 5 industries from the first sample and 10 industries from the second sample cannot affect the results significantly. The 10 omitted industries contributed less than 2 percent of the total value-added in 1977.

REFERENCES

Bell, Frederick W., "The Effect of Monopoly Profits and Wages on Prices and Consumer Surplus," *Western Economics Journal*, June 1968, *6*, 233–41.

Gisser, Micha, "Welfare Implications of Oligopoly in Food Manufacturing," *American Journal of Agricultural Economics*, November 1982, *64*, 616–24.

______, "Price Leadership and Dynamic Aspects of Oligopoly in U.S. Manufacturing," *Journal of Political Economy*, December 1984, *6*, 1035–48.

Harberger, Arnold C., "Monopoly and Resource Allocation, "*American Economic Review Proceedings*, *May* 1954, *44*, 77–87.

______, *Taxation and Welfare*, Boston: Little, Brown, 1974.

Hirshleifer, Jack, *Price Theory and Applica-*

tions, 3rd ed., Englewood Cliffs: Prentice-Hall, 1974.

Kamerschen, David R., "An Estimation of the 'Welfare Losses' from Monopoly in the American Economy," *Western Economic Journal*, Summer 1966, *4*, 221–36.

Kendrick, John W. and Grossman, Elliot S., *Productivity in the United States, Trends and Cycles*, Baltimore: Johns Hopkins University Press, 1980.

Lustgarten, Steven, "Gains and Losses from Concentration: A Comment," *Journal of Law and Economics*, April 1979, *22*, 183–90.

Peltzman, Sam, "The Gains and Losses from Industrial Concentration," *Journal of Law and Economics*, October 1977, *20*, 229–63.

Schwartzman, David, "The Burden of Monopoly," *Journal of Political Economy*, December 1960, *68*, 627–30.

Siegfried, John J. and Tiemann, Thomas K., "The Welfare Cost of Monopoly: An Inter-Industry Analysis," *Economic Inquiry*, June 1974, *12*, 190–202.

Stigler, George J., "A Theory of Oligopoly," *Journal of Political Economy*, February 1964, *72*, 44–61.

Worcester, Dean A., Jr., "New Estimates of the Welfare Loss to Monopoly: U.S. 1956–69," *Southern Economic Journal*, October 1973, *40*, 234–46.

______, "On Monopoly Welfare Losses: Comment," *American Economic Review*, December 1975, *65*, 1015–23.

U.S. Bureau of the Census, *Census of Manufactures*, Washington, 1977.

Child Support, Welfare Dependency, and Poverty

By PHILIP K. ROBINS*

This paper presents an empirical analysis of the effect of child support enforcement policies on poverty and welfare dependency among female-headed families. A special supplement to the April 1982 Current Population Survey provides the data for the analysis. The results suggest that child support enforcement may represent an effective means for reducing welfare program costs, but is unlikely to have a dramatic effect on reducing either poverty or welfare dependency.

Female-headed families have among the highest poverty rates of any major demographic group in the United States. In 1982, 48 percent of all female-headed families were poor, compared with 10 percent of other types of families (U.S. House of Representatives, 1983a). Despite the fact that only one-fifth of all families with children are headed by women, this group constitutes the majority (55 percent) of all poor families.

With an increasing rate of illegitimacy and a high divorce rate, the size of the female-headed population continues to grow.[1] In 1960, there were 1.9 million female-headed families in the United States, or 7 percent of all families. By 1983, the number of female-headed families totaled 5.7 million, or 19 percent of all families (U.S. Department of Health and Human Services, 1983).

The increase in the number of female-headed families has resulted in a growing number of children not living with both natural parents. Several authors, including Daniel Moynihan (1981), Larry Bumpass and Ronald Rindfuss (1979), and Frank Furstenberg et al. (1983) have projected that by the year 2000, as many as one-half of all children born in the United States will not have spent their entire childhood living with both natural parents. If these projections bear out, it is quite possible that the overall incidence of poverty and welfare dependence in the United States will increase significantly unless new policies are developed for increasing the economic well-being of female-headed families.

The traditional approach adopted by policymakers to reduce poverty and welfare dependency in female-headed families has centered on increasing the employment of the mother.[2] While such policies may have had an impact, they have raised important and difficult tradeoffs concerning the well-being of the younger children. Nevertheless, despite the existence of work requirements,

*Professor of Economics, University of Miami, Coral Gables, FL 33124. I acknowledge the helpful comments of Irwin Garfinkel, Robert Moffitt, participants in workshops at the University of Wisconsin, Syracuse University, and the University of Miami, and three anonymous referees. I also thank William Greene for helpful econometric advice and Henry Sniezek and Moises Tacle for dedicated research assistance. A major portion of this research was supported by funds provided to the Institute for Research on Poverty, University of Wisconsin, by the Department of Health and Human Services under the Small Grants for Visitors Program, Grant No. 40A-83. The opinions expressed in this paper are my own and should not be construed as representing the opinions of any government agency.

[1]About 1.2 million divorces occur annually in the United States (compared with about 2.4 million marriages) and the rate of illegitimate births increased from 5.3 percent in 1960 to 18.9 percent in 1981 (U.S. Department of Commerce, 1985).

[2]One such attempt is the Work Incentive (WIN) program which requires Aid to Families with Dependent Children (AFDC) mothers with children over the age of 6 (over the age of 3 under current proposals) to be available for work or training programs. Refusal to accept WIN services can result in lost of AFDC eligibility. Most welfare programs impose some form of work requirement as a condition of eligibility for receipt of benefits.

few women on welfare hold jobs,[3] and the incidence of poverty remains high for this group.

One important alternative (or perhaps complement) to policies aimed at increasing employment of the mother is to collect child support from the absent father. Such an approach has been receiving greater attention in recent years. In 1975, Congress established the Child Support Enforcement Program as Part D of Title IV (IV–D) of the Social Security Act.[4] The IV–D legislation requires each state to develop a child support enforcement program that provides services for establishing paternity, locating absent parents, establishing child support obligations, and enforcing such obligations. The states are required to provide these services to all AFDC families and to non-AFDC families who request such services.[5] To facilitate collection across states, a federal Parent Locator Service was established with access to federal data files on individuals, including Social Security Administration earnings records and Internal Revenue Service tax records.

Very few women on welfare receive child support from the absent father.[6] According to data from the *Current Population Survey* (*CPS*) for 1981, only 15 percent of full-year recipients of AFDC benefits received child support and only 28 percent had a formal child support award.[7] Clearly, there are a large number of absent fathers who are contributing nothing to the support of their children. However, the problems encountered in collecting support are reflected in the fact that almost one-half of the children in AFDC families had parents who were not married,[8] and in 47 percent of the cases, the whereabouts of the father is reported as unknown, although the true percentage of missing fathers is probably less because the implicit 100 percent tax rate on child support income by the AFDC program creates incentives for the mothers to conceal this information.

Despite difficulties in collecting support, the IV–D Program has grown steadily since its inception. In fiscal year 1983, child support collections on behalf of AFDC families totaled $880 million, or about 6.6 percent of AFDC benefits paid (U.S. Department of Health and Human Services, 1983). Collections were over $1.1 billion for non-AFDC families.[9]

[3]According to the 1979 AFDC *Recipient Characteristics Study*, a random survey of welfare caseworkers, only about 15 percent of women receiving AFDC benefits held full- or part-time jobs in the survey month (March). It should be noted that in this paper, welfare dependency is characterized in terms of participation in the AFDC program. It should be understood, however, that other means-tested programs, such as food stamps, supplemental security income, medicaid, and housing assistance, also constitute part of what is generally termed the "welfare system" in the United States.

[4]A detailed discussion of the legislative history of Title IV–D is given in U.S. Department of Health, Education, and Welfare (1976).

[5]In all but two states (Mississippi and South Carolina), child support collections made on behalf of AFDC families are used to offset AFDC benefit amounts on a dollar-for-dollar basis. Because the AFDC mother does not gain financially from receipt of child support (as long as she remains on the welfare rolls), she has no incentive to seek child support or to report any collections to the welfare authorities. Hence, as a condition of eligibility for receipt of benefits, AFDC authorities require AFDC mothers to assign their support rights to the IV–D agency, who in turn pursues collection.

[6]Yoram Weiss and Robert Willis (1985) present an economic model explaining why absent parents fail to pay child support. They argue that children may be viewed as collective consumption goods and that upon divorce, the absent parent loses control over the allocation of family resources and hence, reduces the amount spent on maintenance of the children. This effect is distinct from the disincentive created by the welfare system in providing resources for the children of divorced parents.

[7]Although the *CPS* is known to undercount the number of families receiving AFDC benefits and is suspected to undercount the number of AFDC families receiving child support (because the IV–D agency makes the collections), these figures are corroborated by data from the 1979 AFDC *Recipient Characteristics Study* as well as from other sources. By way of contrast, 45 percent of non-AFDC women received child support in 1981 and 55 percent had a formal child support award. One should not conclude from these figures, however, that lack of child support is the most important factor contributing to welfare dependence.

[8]Never-married mothers constitute about one-quarter of all female-headed families in the United States, according to the April 1982 *CPS*.

[9]Official statistics on child support collections for non-AFDC families should be viewed with caution because they are suspected of being highly inaccurate,

Two important changes have occurred recently in the IV–D Program that are likely to increase collections in the future. First, section 2331 of the Omnibus Reconciliation Act of 1981 (P.L. 97–35) authorizes the Internal Revenue Service (IRS) to withhold federal income tax refunds for persons seriously delinquent in child support payments to AFDC families. The Office of Child Support Enforcement, the federal agency that administers the program, acts as the agent of IRS in the tax-refund intercept process. This is the first time the IRS has participated in a major collection activity not directly related to tax liabilities. It signifies an important new direction in social policy legislation in the United States.[10]

The second important change in the IV–D Program is that it appears to be focusing greater attention on non-AFDC families. The main purpose of the non-AFDC component of the program is "cost avoidance"; that is, preventing families from going on AFDC (and other welfare programs) by collecting child support payments. The Child Support Enforcement Amendments of 1984 (P.L. 98–378), signed into law by President Reagan in August 1984, are designed to aid non-AFDC families in collecting child support. The most important feature of this new legislation is that it authorizes mandatory wage withholding of child support from delinquent parents. The legislation requires employers to withhold child support from paychecks, just as they do for federal and state income taxes, when the child support payments are not being made on a regular basis.

With increased federal and state involvement in child support enforcement, there has emerged a significant need for developing ways of evaluating the government's role in this area. The objective of this paper is to investigate empirically the effectiveness of current child support policies and to determine their role in reducing welfare dependency and poverty. To accomplish this objective, I first develop and estimate a model that shows how receipt of child support influences the welfare participation decision. I then specify a regression model in which the effects of the child support enforcement program on receipt of child support are estimated. Because the data used to estimate program effects are not experimental, particular attention is paid to various forms of self-selection bias that can invalidate the results. Finally, the results are used to simulate the effects of various child support policies on welfare dependency, poverty, and welfare costs. The data used to investigate these issues are from a special supplement to the April 1982 *CPS* in which detailed information about child support and welfare receipt was collected from a nationally representative sample of women with children.

The remainder of this paper is organized as follows. In Section I, the theoretical and empirical model of AFDC participation is presented. In Section II, the results of the AFDC participation model are used in conjunction with data on the Child Support Enforcement Program to derive estimates of the effect of the program (as it existed in 1981) on receipt of child support. In Section III, the empirical findings are used to simulate the effects of various child support enforcement policies on welfare dependency, poverty, and welfare costs. Finally, in Section IV, the main conclusions are summarized.

I. A Model of Welfare Dependency

A. *Theoretical Framework*[11]

A female-headed family is assumed to participate in the AFDC program if par-

possibly overstating true collections by as much as several hundred million dollars (U.S. Department of Health and Human Services, 1982b).

[10] States are also required to have similar programs for state tax refunds, and laws exist in several states to withhold Unemployment Insurance benefits from delinquent absent parents.

[11] The model presented in this section is similar in many respects to models analyzed by myself and Richard West (1980), Orley Ashenfelter (1983), and Robert Moffitt (1983). However, unlike the model analyzed by Moffitt, there is no attempt to explicitly parameterize

ticipation increases its utility. Consider a family that is hypothesized to maximize a monotonic, strictly quasi-concave utility function $U(H, Y)$, where H is hours of work, Y is expenditures on market goods, $U_H < 0$, $U_Y > 0$, $U_{HH} < 0$, $U_{YY} < 0$, and $U_{HY} < 0$. The budget constraint for the family is $Y = WH + N + PB$, where W is the wage rate, N is nonwage income other than AFDC (primarily child support), P is a binary (1,0) variable indicating whether or not the family receives AFDC benefits, and B is the level of AFDC benefits. The AFDC benefit formula may be written as $B = G - N - tWH$, where G is the AFDC guarantee level (benefit when all other sources of income are zero) and t is the implicit AFDC tax rate.[12] Using the AFDC benefit formula, the budget constraint can be rewritten as $Y = WH(1 - Pt) + N(1 - P) + PG$.

Maximization of the utility function subject to the budget constraint yields a set of equations determining Y, H, and the marginal utility of income as functions of W, P, t, N and G. Substituting these solution equations into the direct utility function yields the indirect utility function $V = V(W(1 - Pt), N(1 - P) + PG)$.

Denoting V^1 as the value of V for AFDC recipients and V^0 as the value of V for nonrecipients, yields the AFDC participation decision: participate in AFDC if $\Delta V = V^1 - V^0 > 0$, do not participate otherwise.

Using this decision framework, it is evident that increases in child support reduce the probability of participating in AFDC. This is because increases in N raise V^0 but not V^1, since N does not appear in V^1. Hence, increases in child support reduce the AFDC break-even point (given by $H_b = (G - N)/Wt$) and reduce the likelihood the family will be eligible for the program.[13]

A second-order Taylor-series expansion of ΔV around the (unobserved) nonrecipient equilibrium position (V^0) yields

$$\text{(1)}\quad \Delta V = V_N(G - N) + V_W(-Wt) + 1/2V_{NN}(G - N)^2 + 1/2V_{WW}(Wt)^2 + V_{NW}(G - N)(-Wt) + \text{remainder},$$

where $G - N$ is the change in net nonwage income and $-Wt$ is the change in the net wage rate induced by the AFDC program.

Assuming an upward-sloping labor supply function, it is expected that $V_N > 0$, $V_W > 0$, $V_{NN} > 0$, $V_{WW} > 0$, and $V_{NW} < 0$. In reality, all of these partial derivatives are likely to vary across families, but in the empirical work below only their average values in the sample are estimated.[14]

This specification provides a convenient framework for calculation of income and substitution effects. This can be seen by making use of Roy's identity, which is given by $H_0 = V_W/V_N$ (see Eugene Silberberg, 1978, or James Henderson and Richard Quandt, 1980). Differentiating Roy's identity with respect to N and W, and using the Slutsky equation gives the income effect,

$$\text{(2)}\quad \partial H_0/\partial N = (V_N V_{WN} - V_W V_{NN})/V_N^2 = (V_{WN} - H_0 V_{NN})/V_N,$$

welfare stigma and no attempt to impose a particular functional form on the utility function. For other recent models of welfare participation, see Mark Plant (1984) and Rebecca Blank (1985).

[12] This is a simplification of the actual AFDC benefit formula used in practice. The exact formula is given in my 1984 paper. Also, note that this specification abstracts from the positive tax system and other tax and transfer programs that are income conditioned.

[13] For certain families, $G - N$ may be actually negative and a break-even point will not exist. These are primarily families with high child support payments who reside in a low guarantee state. Nonwage income N does not appear in V^0 because of the assumption that it is taxed at the rate of 100 percent by the AFDC program (which, as indicated earlier, is true in all but two states).

[14] For example, if a linear labor supply function is specified ($H = a + bW + cN$), then $V_w = V^2c^2e^{cW}H$, $V_N = V^2c^2e^{cW}$, $V_{WW} = Vc^2e^{cW}(Vb + 2HV_W + VHc)$, $V_{NN} = 2c^2e^{cW}VV_N$, and $V_{WN} = Vc^2e^{cW}(Vc + 2V_W)$, where $V = (Se^I)^{-1}$, S being the substitution effect $(b - cH)$ and I the total income elasticity (cW). This specification has been used by Moffitt, among others.

and the substitution effect,

$$(3)\quad \partial H_0/\partial W|_u = \left(V_{WW} - 2H_0V_{NW} + H_0^2V_{NN}\right)/V_N.$$

Hence, estimates of the five partial derivatives of (1) provide enough information to calculate income and substitution effects at the nonrecipient equilibrium position.

The first- and second-order terms in (1) may be conveniently interpreted in terms of mechanical and behavioral effects on participation (see Orley Ashenfelter). Mechanical effects determine participation solely through income eligibility for the program (i.e., in the absence of any labor supply effects caused by changes in the net wage rate and net nonwage income). Behavioral effects determine participation through changes in labor supply that make an otherwise ineligible person eligible for benefits. Using Roy's identity and ignoring the remainder term, (1) can be rewritten as

$$(4)\quad \Delta V = V_N\Bigg[B + 1/2(Wt)^2 \times\left(\left(V_{WW} - 2V_{NW}\left(\frac{G-N}{Wt}\right) + V_{NN}\left(\frac{G-N}{Wt}\right)^2\right)\Big/V_N\right)\Bigg],$$

where $B = G - N - tWH_0$ is the AFDC benefit evaluated at the nonrecipient equilibrium position.

Equation (4) decomposes the participation decision into these two components. The first component, B, indicates that participation occurs if the AFDC benefit, calculated at the nonrecipient position, is positive. The second component, which is necessarily positive so long as $G > N$, indicates that participation may occur even if the person is initially ineligible for benefits (above the break-even level). Using equation (3) along with the fact that $H_b = (G-N)/Wt$, (4) can be rewritten as

$$(5)\quad \Delta V = V_N\left[B + 1/2(Wt)^2 S_b\right],$$

where S_b is the substitution effect *at the break-even level.* Equation (5) is the utility equivalent of the excess expenditure function analyzed by Ashenfelter, where V_N is the appropriate conversion factor.

At the break-even level, (5) says that the person participates if $S_b > 0$ (i.e., if the indifference curves are not right angles). Above the break-even level ($B < 0$), the person participates if $S_b > 2|B|/(Wt)^2$, assuming rational behavior and no stigma.[15]

This discussion suggests two different formulations that can be tested empirically; equation (1) which yields estimates of income and substitution effects at the nonrecipient position and equation (5) which yields an estimate of the substitution effect at the break-even level. As indicated in the next section, a third formulation is also tested which constrains the uncompensated wage effect at the nonrecipient position to be zero. These three estimation procedures represent different assumptions about which partial derivatives are treated as parameters for purposes of estimation. In equation (1), H_0, V_N, V_W, V_{NN}, V_{WW}, and V_{NW} are treated as parameters to be estimated. In equation (5), V_N and S_b are treated as parameters to be estimated. In the third formulation, V_W, V_N, and V_{NN} are treated as parameters to be estimated and V_{WW} and V_{NW} are assumed to be zero. All of these estimates can be used to assess the relative importance of mechanical and behavioral effects on the AFDC participation decision.

B. *Estimation*

If the remainder in (1) is assumed to be normally distributed with unit variance, then (1) becomes simply a probit model. Because AFDC is a monthly program, proper estimation of (1) requires monthly data on AFDC participation. Unfortunately, the data file used to estimate the model (the 1982 *CPS*) does not contain monthly data on AFDC

[15]As indicated earlier, Moffitt develops a model of AFDC participation in which there may be stigma associated with participation. In his model, participation may not always occur when $B > 0$.

participation. However, the *CPS* does contain information sufficient for estimating a model determining the number of months on AFDC during the survey year, or equivalently, the fraction of the year spent on AFDC (number of months divided by 12).[16] Thus empirically, the focus is on the AFDC participation decision over a 12-month period rather than in any given month. The model actually estimated has the general form of a two-limit probit (Tobit) regression model (see Richard Rosett and Forrest Nelson, 1975).[17]

C. *Data and Variables*

The data, as indicated earlier, are from a special supplement to the April 1982 *Current Population Survey*. This survey is the second attempt by the Department of Commerce to obtain detailed information about child support arrangements of families in which the children are not living with both natural parents.[18] Although the survey covers families in which the mother is currently either married or unmarried, the focus in this paper is on families with only one parent in the home because these families are at greater risk of becoming dependent on welfare. The 1982 *CPS* is particularly useful for this study because it contains information on participation in the AFDC program as well as on various services performed by administrators of the Child Support Enforcement Program on behalf of AFDC and non-AFDC families.[19]

The main variables included in the empirical participation model are those contained in equation (1), namely $G-N$ and Wt.[20] In addition, some control variables are

[16]The *CPS* does not contain information on whether the months of AFDC participation coincide with receipt of earnings, child support, or other income, which is a significant weakness for studying transfer programs that use a monthly accounting period. The newly developed *Survey of Income and Program Participation*, which was not available at the time this study was undertaken, should rectify these difficulties because it collects transfer program data on a monthly basis.

[17]For families receiving AFDC benefits for only part of the year, the *CPS* does not identify the number of months benefits were received, only that they were received for part of the year. As Rosett and Nelson show, this information is sufficient to identify all the parameters of the model. The model distinguishes three groups of individuals: full-year recipients, partial-year recipients, and nonrecipients. The log likelihood function is $L = \log[\text{Prob}(F=0)] + \log[\text{Prob}(0<F<1)] + \log[\text{Prob}(F=1)]$, where F is the fraction of the year participating in AFDC.

[18]The first survey took place in April 1979 and the most recent survey took place in April 1986. There are also plans to continue the survey on a regular basis. Each child support survey has been merged (by the U.S. Census Bureau) with the March *CPS* of that year. Thus, in addition to child support information, the public use files contain a considerable amount of economic and demographic information for each family. See U.S. Department of Commerce (1981, 1983) for a description of the first two surveys.

[19]As indicated earlier, although the *CPS* is known to understate the number of families receiving AFDC benefits, it is not known whether the unidentified families are a random subset of the *CPS* population. If they are a random subset, the estimates presented in this paper will not be biased.

[20]Because gross wage rates are not observed for nonworkers, wages for the entire sample were imputed using the James Heckman (1979) two-step method. The coefficients used in the imputations are presented in my 1984 paper. The imputations are not conditioned on actual work status (the unconditional expectations are used). Variables in the wage equation include dummy variables for geographic region, city size, race/ethnicity, and being a high school graduate, and quadratic terms for education and work experience. The probit selection equation contains the same variables plus dummy variables for headship status, homeownership, marital status, and the amount of nonwage income received in 1981. The selectivity bias correction term is not statistically significant in the wage equation, consistent with previous findings for women. Tax rates used in calculating t were obtained using a modified version of the procedure proposed by Robert Hutchens (1978). Details are given in my 1984 paper. Prior to 1982, t varied across states because of the deductibility of work expenses which increased with earnings. Hence, the AFDC tax rate consisted of the statutory rate (.67 in 1981) less the increase in work expenses per dollar increase in earnings. In 1982, the AFDC regulations changed when work expenses were standardized and the basic tax rate was increased to 100 percent (after 4 months of earnings). These changes were implemented after the time period covered in this study. In our sample, the mean W is \$4.58 and the mean t is .39. Guarantee levels used for calculating G were obtain from U.S. Department of Health and Human Services (1981, 1982). These guarantee levels varied both across states and by family size. In 1981, for a family of 4, G ranged from a low of \$122 in Mississippi to a high of \$560 in Vermont. Other

TABLE 1—ESTIMATES OF BASIC AFDC PARTICIPATION MODEL
(No Behavioral Effects)

Independent Variables	Mean	Normalized Coefficient	Asymptotic Standard Error
$-Wt$	-1.72	.151[a]	.079
$(G-N)(\times 10^3)$	2.66	.152[b]	.016
1 = Northeast	.20	.273[b]	.095
1 = Northcentral	.24	.361[b]	.088
1 = West	.22	.087	.098
1 = Black	.31	.406[b]	.072
1 = Hispanic	.08	−.027	.111
Years of Schooling	11.80	−.110[b]	.015
1 = Divorced	.49	−.037	.089
1 = Separated	.25	.018	.087
1 = Worked Full Time at Time of Dissolution	.31	−.665[b]	.078
1 = Worked Part Time at Time of Dissolution	.07	−.362[b]	.124
1 = Unemployed at Time of Dissolution	.08	−.008	.105
Family Size	3.15	.013	.023
Age of Mother	32.45	−.027[b]	.004
1 = Limit on State AFDC Payment or Ratable on Deficit	.31	−.050	.066
Constant	–	1.394[b]	.207
Reciprocal of Residual Standard Error	–	.306[b]	.020
Sample Statistics			
− Log of Likelihood		1,709	
Sample Size		2,543	
Nonrecipients		1,692	
Partial-Year Recipients		206	
Full-Year Recipients		645	

Notes: The variable W is the gross wage rate, t is the AFDC tax rate, G is the AFDC guarantee level, and N is nonwage income other than AFDC (primarily child support).
[a]Statistically significant at the 10 percent level.
[b]Statistically significant at the 1 percent level.

added to account for varying preference structures of families.[21] These include dummy variables for region of the country (Northeast, Northcentral, West), dummy variables for race/ethnicity (black, Hispanic), age and years of schooling of the mother, family size, dummy variables for marital status of the mother (divorced, separated), dummy variables for employment status of the mother at the time of the marital dissolution (working full time, working part time, and unemployed,[22] and a dummy vari-

nonwage income used to calculate N (including child support) was taken directly from the *CPS*.

[21]A more general specification (not pursued here) would be to allow for varying preferences through interaction of the control variables with the basic budget constraint variables in (1) and to consider explicitly the interaction of the AFDC system with other tax and transfer programs.

[22]Employment status of the mother at the time of the marital dissolution is a proxy for her preference for work. Because of the way the survey was designed, this variable is available only for women who were previously married. Because variables are included for marital status, the omitted category (never married) will pick up effects of marital status and work preference on AFDC participation.

TABLE 2—ESTIMATES OF BEHAVIORAL EFFECTS ON PARTICIPATION

		Normalized Coefficient		
Variable	Mean	(1)	(2)	(3)
$-Wt$	-1.72	.187	.075	.145[a]
		(.278)	(.270)	(.079)
$(G-N)(\times 10^3)$	2.66	.027	.152[c]	.127[c]
		(.051)	(.016)	(.019)
$(Wt)^2$	3.29	$-.038$	$-.022$	–
		(.078)	(.075)	
$(G-N)^2(\times 10^6)$	14.27	.006[a]	–	.006[b]
		(.003)		(.003)
$(G-N)(-Wt)(\times 10^3)$	-4.38	$-.061$[b]	–	–
		(.029)		
$-$Log of Likelihood		1,705	1,709	1,707

Notes: See Table 1. Coefficients of control variables are not reported. Asymptotic standard errors are shown in parentheses.
[a]Statistically significant at the 10 percent level.
[b]Statistically significant at the 5 percent level.
[c]Statistically significant at the 1 percent level.

able indicating whether the state in which the family resides imposes a limit on the AFDC payment or imposes less than a 100 percent tax rate on child support payments and other nonwage income.

D. *Results*

Estimates of the basic AFDC participation model (no behavioral effects) are presented in Table 1, along with the means of the variables. The two estimated first-order partial derivatives of the indirect utility function (V_W and V_N) are statistically significant and of the expected sign. The results suggest that differences across states in eligibility requirements (as reflected in guarantee levels and tax rates) play a significant role in determining AFDC participation. It is important to note these differences are not due to labor supply responses to the program's guarantee levels and tax rates, but rather to differences in program generosity across states (as measured by the break-even level). In states with higher guarantee levels and lower tax rates (higher break-even levels) eligibility for the program (and hence participation) is more likely. The results also imply that participation is higher for low-wage women and lower for women that receive child support payments, again because of break-even differences only. The implied nonrecipient annual hours of work from the first-order terms is 992, which is very close to average observed hours of work in the sample of 1,101. The standard deviation of observed hours of work is 951.

The results also indicate that the mother's preference for work is a strong determinant of AFDC participation. Mothers who were employed at the time of the marital dissolution are much less likely to become welfare recipients than mothers who did not work. The effect of full-time work is almost twice the effect of part-time work. Mothers who were unemployed at the time of the marital dissolution are just as likely to become welfare recipients as mothers who were out of the labor force.

Table 2 presents estimates of three versions of the model that allow for behavioral effects in addition to mechanical effects. The first version is equation (1). The second version is equation (5). The third version is a hybrid which constrains the uncompensated wage effect at the nonrecipient position to be zero. This third model is estimated because, as will be argued below, collinearity between Wt and $(Wt)^2$ makes it difficult to produce reliable estimates of V_W and V_{WW}. For brev-

ity, the coefficients of the control variables are not reported (they are similar to the estimates in Table 1).

In the more general formulation in which all five first- and second-order terms are estimated (col. 1, Table 2), only two are statistically significant, although four are of the correct sign. A likelihood ratio test indicates that the second-order terms are statistically significant at the 5 percent level (*Chi*-square = 8.16, degrees of freedom = 3); hence behavioral effects appear to be operating. Nevertheless, the implied income and substitution effects (derived using equations (2) and (3)) are inordinately large compared to generally accepted values.[23] In the version of the model in which V_N and S_b are treated as parameters, (col. 2), the estimated substitution effect at the break-even level is small and negative, but is not statistically significant.[24] Hence, this version of the model provides no evidence of behavioral effects on participation, in the sense that an otherwise ineligible woman reduces her labor supply in order to become eligible for benefits.[25]

The absence of precisely estimated behavioral effects in columns 1 and 2 of Table 2 may be due to the high collinearity between the linear and quadratic net wage terms (which may have arisen because of lack of sufficient variation in predicted wages and predicted AFDC tax rates). In column 3, estimates are presented of a model in which only the linear wage term and linear and quadratic terms in the net guarantee variable appears. This version of the model constrains the uncompensated wage effect to be zero, which does not appear to be an unreasonable assumption for female heads of families.[26] In this restricted version of the model, all three coefficients are statistically significant. The implied nonrecipient hours of work is 1,140 which is very close to average observed hours in the sample. The implied income effect (per thousand dollars) is -113.5 and the implied substitution effect is 129.4. Both estimates are similar to those reported in my earlier article (1985, Table 8, p. 580) and are well within the range reported in other studies. The likelihood ratio test for the quadratic term (*Chi*-square = 3.92, degrees of freedom = 1) is significant at the 5 percent level.

Taken together, the results in Table 2 provide some evidence of a behavioral effect on AFDC participation. However, the implausible estimates for specifications (1) and (2) are somewhat troublesome and imply the results should be viewed with caution.

II. Effectiveness of Current Child Support Enforcement Procedures

In this section, the *CPS* data are used to estimate the effect of the IV–D Program on various child support outcomes. Because program effects are likely to vary by AFDC status, separate estimates are derived for AFDC and non-AFDC families.[27] As indicated below, the results from the AFDC participation model just presented play an important role in the analysis of this section.

[23] Because the estimated V_W is almost seven times the estimated V_N, the implied income and substitution effects are roughly seven times larger than coventionally accepted estimates.

[24] Ashenfelter also obtained an insignificantly estimated substitution effect.

[25] The absence of a behavioral effect at the break-even level does not necessarily imply the absence of a labor supply effect for persons below the break-even level, because the substitution effect may vary with labor supply.

[26] My 1985 article reports close to a zero uncompensated wage effect for female heads of households, based on data from the income maintenance experiments. However, Moffitt finds a significant uncompensated wage effect for female heads of families using data from the *Michigan Panel Study of Income Dynamics*.

[27] The AFDC families are required to assign their support rights to the state IV–D agency while participation in the IV–D Program is strictly voluntary for non-AFDC families. Furthermore, AFDC families do not gain financially from collection of child support while non-AFDC families do. For these and other reasons, the effectiveness of child support enforcement procedures may be different for the two groups of families.

A. *Empirical Specification*

The basic empirical model of the effects of current child support procedures is given as follows:

$$C_i = Z_i\gamma_i + IVD_i\delta_i + u_i, \tag{6}$$

where C_i = the child support outcome for the ith group, Z_i = a vector of control variables for the ith group, IVD_i = a vector of variables representing services provided by the IV–D Program for the ith group, and u_i = a random error term. Three distinct groups are identified for this analysis: nonrecipients of AFDC, partial-year recipients of AFDC, and full-year recipients of AFDC. The effect of the IV–D Program on child support for each group is given by the estimate of δ_i.

In empirically implementing equation (6), three potentially serious problems arise. First, if selection into each of the three groups depends on unmeasured variables affecting AFDC status, then standard regression analysis applied to (6) will yield biased estimates of δ_i. In general, the size of the bias will depend on the degree of correlation between the error term in (6) and the error term in the AFDC participation equation.

Extending the analysis of James Heckman (1979) to the multiple selection case, selectivity bias correction terms are constructed for each group based on the estimates of the AFDC participation model. The general form of the selectivity correction term is $\lambda = (f_1 - f_u)/(F_u - F_1)$, where f is the normal density function, F is the corresponding distribution function, and the 1 and the u subscripts refer to the lower and upper truncation points, respectively. In the case of nonrecipients of AFDC, $1 = -\infty$ and $u = 0$, so that $\lambda = -f_0/F_0$. In the case of full-year recipients of AFDC, $1 = 1$ and $u = \infty$, so that $\lambda = f_1/(1 - F_1)$. In the case of part-year recipients of AFDC, $1 = 0$ and $u = 1$ so that $\lambda = (f_0 - f_1)/(F_1 - F_0)$. These selectivity correction terms are entered into the appropriate equations and standard regression analysis is applied.[28] The coefficients of the selectivity terms are used to derive the correlation coefficient between the error term in (6) and the error term in the AFDC participation equation.

The second and third problems arising in estimating (6) stem from the fact that IV–D services are not provided on a random basis to the population. The second problem has to do with the fact that not all families seek help from the child support agency. In particular, only families having difficulties obtaining child support are likely to apply. This type of selectivity bias is particularly relevant for non-AFDC families because their participation in the program is voluntary. Because all AFDC families are required to assign support rights to the child support agency, selection into the program may not cause as serious a bias for them, although such a bias still may be present. Failure to correct for this problem could lead to a significant underestimate of the impact of the Child Support Enforcement Program.

The third problem has to do with the possibility that among those who apply for services, the child support agency targets services in a nonrandom way. This potential source of bias is particularly relevant for AFDC families because program services are provided to them free of charge.[29] Faced with resource constraints and performance standards, there may be an inducement to "cream," that is, provide services to the easiest cases. If such creaming exists, estimated impacts of the IV–D Program would be too high.

In the empirical work below, an attempt is made to adjust for nonrandom provision of services. For the problem of nonrandom selection into the IV–D Program, the adjustment is made by including as a control variable a dummy variable denoting whether the

[28]The estimated standard errors of the coefficients are corrected for bias due to heteroskedasticity using the estimator developed by William Greene (1981).

[29]States must charge a fee to non-AFDC families and some have cost recovery provisions, so the incentive to target services for them is not as great.

family reported having contacted the child support agency. The expected sign of this variable is negative, particularly for non-AFDC families.

The problem of nonrandom targeting is more difficult to deal with empirically. It essentially requires purging the program service variables of systematic unmeasured effects. Since this problem is likely to be most important for AFDC families, an adjustment is attempted only for them.[30]

B. *Variables*

The *CPS* provides information on the types of services provided by the IV–D Program. Mothers were asked whether they had ever contacted the child support agency, whether they received help, and what types of services were provided. The services listed are 1) an attempt to locate the father, 2) an attempt to establish paternity, 3) an attempt to establish a support obligation, 4) an attempt to enforce a support order, 5) an attempt to obtain collection, and 6) other (unspecified) services. Three sets of variables measuring program effects are constructed from this information and each is included in a separate regression. The first set consists of a single dummy variable indicating whether any help was received. The second set consists of a single variable denoting the number of different types of services provided. The third set consists of six dummy variables denoting which specific services were provided. Four child support outcomes are examined as dependent variables: whether child support was received, the amount of child support received, whether a child support obligation exists, and the amount of the child support obligation.[31]

In addition to the program effect variables, the empirical models contain several control variables. The control variables consist of dummy variables for region of the country (Northeast, Northcentral, West), dummy variables for race/ethnicity (black, Spanish), age and education of the mother, number of children in various age groups (0–5, 6–11, 12–18), dummy variables for marital status of mother (divorced, separated), years since the marital dissolution, number of child support enforcement procedures used in the state (see U.S. Department of Health and Human Services, 1982a), a dummy variable for whether the state has a tax intercept program, a dummy variable for whether the state has a statute of limitation for establishing paternity, and the two selectivity correction terms discussed above.[32]

C. *Program Effects on Receipt of Child Support*

Table 3 presents for the three different specifications of program services the estimated effects of the IV–D Program on

[30]A generalized least squares, instrumental variable procedure was adopted (to account for the endogeneity of the program service variable) and the results are reported in my 1984 paper. The results imply that IV–D Program administrators appear to target services on the more difficult rather than the easier cases. Hence, the data provide no evidence of "creaming" by program administrators. The main results presented below do not adjust for nonrandom targeting so, if anything, they underestimate the true effect of the IV–D Program.

[31]Of the sample, 36 percent reported receiving child support in 1981 and 48 percent reported having a legal obligation. These percentages are lower for AFDC recipients and higher for nonrecipients. Among those receiving child support, the average amount received was $2,071 in 1981 and among those with an obligation, the average amount of the obligation was $2,395. Of the sample, 26 percent reported that they had at some time contacted the child support agency while 13 percent reported actually receiving services. The program service variables are somewhat unreliable for AFDC families because the family may not be aware that services are being provided for them. In my 1984 paper, I utilize a caseworker sample (the 1979 AFDC *Recipient Characteristics Study*) to estimate program effects for AFDC families. Although program utilization rates are greater in the caseworker sample, the estimated program effects are very similar to those obtained using the *CPS* sample.

[32]Several of the variables in the program effect equations are excluded from the AFDC participation equation and vice versa. Hence identification of the program effect equations is achieved through these exclusion restrictions as well as from the nonlinearity of the selectivity correction terms.

TABLE 3—ESTIMATED EFFECTS OF CHILD SUPPORT ENFORCEMENT PROGRAM ON RECEIPT OF CHILD SUPPORT

	Independent Variables							
	(1)	(2)	(3)					
Dependent Variable	1 = Received Help From OCSE	Number of Services Provided	1 = Attempt to Locate Father	1 = Attempt to Establish Paternity	1 = Attempt to Establish Obligation	1 = Attempt to Enforce Obligation	1 = Attempt to Obtain Collection	1 = Other Service
Probability of Receiving Child Support								
Full Sample	.20[c]	.09[c]	−.14[c]	−.01	.12[c]	.22[c]	.21[c]	−.02
($N = 2{,}543$)	(.03)	(.02)	(.05)	(.09)	(.05)	(.05)	(.05)	(.05)
Nonrecipients of AFDC	.19[c]	.09[c]	−.10	−.10	.06	.22[c]	.19[c]	.02
($N = 1{,}692$)	(.05)	(.03)	(.07)	(.20)	(.08)	(.06)	(.07)	(.08)
Partial-Year Recipients	.29[c]	.11[b]	−.22[a]	.10	.13	.11	.37[b]	.21
of AFDC ($N = 206$)	(.10)	(.05)	(.14)	(.21)	(.13)	(.14)	(.17)	(.15)
Full-Year Recipients	.18[c]	.08[c]	−.15[c]	−.10	.18[c]	.16[a]	.26[c]	−.02
of AFDC ($N = 645$)	(.04)	(.02)	(.05)	(.09)	(.05)	(.06)	(.06)	(.06)
Amount Received of Child Support								
Full Sample	258.4[c]	113.4[b]	−188.8	290.4	319.0[b]	111.7	180.9	−175.5
($N = 2{,}543$)	(100.6)	(52.0)	(143.0)	(292.9)	(146.4)	(146.4)	(155.9)	(166.9)
Nonrecipients of AFDC	208.0	120.6	−124.1	506.1	350.7	86.1	191.3	−109.0
($N = 1{,}692$)	(173.2)	(92.8)	(271.3)	(720.4)	(276.1)	(236.1)	(266.7)	(305.5)
Partial-Year Recipients	284.6	42.4	−208.3	62.7	177.1	241.6	−82.3	32.3
of AFDC ($N = 206$)	(191.7)	(97.7)	(271.6)	(422.4)	(266.7)	(275.1)	(330.7)	(301.6)
Full-Year Recipients	293.1[b]	122.3[b]	−247.3[b]	27.8	238.1	166.3	485.0[b]	−32.3
of AFDC ($N = 645$)	(116.5)	(56.8)	(149.5)	(265.0)	(149.1)	(195.4)	(181.8)	(185.9)

Notes: All equations contain also a set of control variables that are described in the text. The estimated effects are for 1981. Standard errors are shown in parentheses.

[a]Statistically significant at the 10 percent level.

[b]Statistically significant at the 5 percent level.

[c]Statistically significant at the 1 percent level.

whether child support is received and the amount received.[33] These results indicate that the program has a significantly positive effect on receiving child support for each group.

In specification 1 for the full sample, families who reported receiving help from the child support agency have a 20 percentage point higher probability of receiving child support than families who did not report receiving such help. Prior to receiving services, the mean probability of receiving child support for those who contacted the agency, is .18.[34] Provision of program services raises this probability to .38. Interestingly, the mean child support recipiency rate for those who did not contact the child support agency is also .38, so the program

[33]These equations are all estimated using ordinary least squares (*OLS*). Technically, more appropriate procedures would be a probit model for the equation determining whether child support is received and a Tobit model for the equation determining the amount of child support received. However, these nonlinear procedures complicate the selectivity corrections somewhat and increase the computational burden. Use of *OLS* does not bias the coefficients (properly interpreted) and should be a reasonable approximation. Table 3 does not report the coefficients of the control variables, for the sake of brevity. The full regression results are available on request from the author.

[34]This figure is derived as $\overline{C} + b$, where $\overline{C}$ is the mean probability for families who have not contacted the child support agency and b is the coefficient on the contact variable, which is $-.20$.

appears to be raising the probability of receiving child support to the level prevailing in the rest of the population. Overall, as of 1981, the program appears to have increased the child support recipiency rate among single-parent families by .027, which is about an 8 percent increase.[35]

Dollarwise, families using the IV–D Program are not receiving as much as the rest of the population. Prior to receiving services, the mean amount of child support received per year for those who contacted the agency is $271. Provision of services increases average collections to $530 per year. The mean amount of child support received by those who did not contact the agency is $874 per year. Overall, the program appears to be increasing collections nationwide by about 5 percent.

The number of services provided has a significant effect on receipt of child support, as indicated from the results of estimating specification 2 of the model (for the full sample). Each service, on average, increases the probability of receiving child support by about 9 percentage points.[36] However, the results from specification 3 suggest that all services do not have the same impact.

The results from specification 3 suggest that successful collection of child support may be the result of a cumulative package of services. Use of the parent locator service to find the absent father, for example, is estimated to actually reduce the probability of obtaining support unless enforcement services are also provided.[37] Similarly, establishing paternity is estimated to have no effect on collection rates unless further services are provided. There appears to be a positive payoff from attempts to establish support obligations, but the results imply that collection rates can be increased even further if enforcement services are also provided.

The program appears to be effective for both AFDC and non-AFDC families. However, in terms of dollars collected, the program appears to be more effective for AFDC families. This may be a reflection of the fact that collections for AFDC families offset AFDC benefit amounts (and hence AFDC program costs) on a dollar-for-dollar basis, so the incentive for program administrators to pursue collection for AFDC families is greater. Collection for non-AFDC families also reduces AFDC program costs, but this "cost avoidance" impact of the program is indirect, resulting from its effect on reducing welfare dependency.[38]

Overall, the results in Table 3 indicate that the program is fairly successful in its enforcement operations. As more comprehensive collection mechanisms are put into effect, such as those contained in recent legislation, program effectiveness is likely to become even greater.

D. *Effect of Selection Bias on the Results*

As indicated earlier, the empirical results are adjusted for two types of selectivity bias; one arising from selection into the AFDC program and one arising from selection into the IV–D Program. Table 4 shows the effects of these two types of selectivity bias on the

[35] This is obtained by multiplying the program effect coefficient from specification (1) (.20) by the fraction of families receiving IV–D services (.134).

[36] The average number of services received by those receiving services is about 1.4.

[37] It is somewhat difficult to believe that use of the parent-locator service actually reduces the probability of obtaining support. The negative sign may indicate that some selectivity bias still remains. However, it is possible the father becomes somewhat hostile when contacted by the child support agency and becomes less likely to pay support unless enforcement actions are also taken. In general, because the program service variables are not interacted with one another, conclusions regarding combinations of services should be viewed as speculative only.

[38] In my 1984 paper, estimates are reported of program effects by marital status of the mother. As would be expected, the effects are largest for divorced women and smallest for never-married women. The effects are smallest for never-married women because usually more services have to be provided for them, the fathers are less willing to pay, and the fathers are less able to pay. Because almost one-half the AFDC caseload consists of women not married to the children's father, this represents a major stumbling block for the IV–D Program in its attempt to reduce welfare dependency through child support collections.

TABLE 4—EFFECTS OF SELECTIVITY CORRECTIONS ON RESULTS

AFDC Status	Estimated Effect of Receiving Help from OCSE	Correction for Selection into the AFDC Program	ρ	Correction for Selection into the IV–D Program
Nonrecipients	208.0	8,211.1[c]	.57	−527.3[c]
(N=1,692)	(173.2)	(597.3)		(129.6)
	285.3[a]	–	–	−657.9[c]
	(155.5)			(117.5)
	−287.6[b]	–	–	–
	(118.2)			
Partial-Year Recipients	284.6	4,106.0[c]	.51	−399.1[b]
(N=206)	(191.7)	(1,164.3)		(171.1)
	314.3	–	–	−526.0[c]
	(198.7)			(177.0)
	−85.2	–	–	–
	(149.3)			
Full-Year Recipients	293.1[b]	5,055.9[c]	.65	−264.4[c]
(N=645)	(116.5)	(366.9)		(98.5)
	264.9[b]	–	–	−327.0[c]
	(108.4)			(90.7)
	16.0	–	–	–
	(84.3)			

Notes: Results are for equations in which the dependent variable is the amount of child support received in 1981. All equations also contain a set of control variables described in the text. Standard errors are shown in parentheses.

[a]Statistically significant at the 10 percent level.
[b]Statistically significant at the 5 percent level.
[c]Statistically significant at the 1 percent level.

results.[39] Of the two, selection into the IV–D Program is clearly the most important.

Selection into the AFDC program causes the effects of the IV–D Program for full-year AFDC families to be underestimated and the effects for part-year and non-AFDC families to be overestimated. For each AFDC status, the selectivity correction term is statistically significant. An estimate of the correlation coefficient between the error term in the child support recipiency equation and the error term in the AFDC participation equation ρ is also given in the table. These estimates imply that the error terms are positively correlated.

Selection into the IV–D Program causes the effects of the IV–D Program to be severely underestimated for each group. In fact, without this correction term (the contact variable), the estimated IV–D effects are negative for two of the groups. The term measuring the bias (the coefficient of the contact variable) is greatest for non-AFDC families (as expected), but is large and statistically significant for AFDC families as well.

E. *Program Effects on Establishing Child Support Obligations*

Collection activities represent an important part of the Child Support Enforcement Program and as indicated earlier, much recent legislation is aimed at improving the collection process. However, collection is only part of the overall child support problem in the United States. Equally important from a policy perspective is the establishment of formal child support obligations. More than one-half the mothers in the *CPS* sample do not have a formal child support

[39]The table shows the effects of selectivity bias for specification 1 in which the dependent variable is the amount of child support received in 1981. The effects of selectivity bias on the other specifications are similar.

TABLE 5—ESTIMATED EFFECT OF CHILD SUPPORT ENFORCEMENT PROGRAM ON HAVING A CHILD SUPPORT OBLIGATION

	Independent Variables			
	(1)	(2)		
Dependent Variables	1 = Received Help From OSCE	1 = Attempt to Locate Father	1 = Attempt to Establish Paternity	1 = Attempt to Establish Obligation
Probability of Having an Obligation				
Full Sample (N = 2,543)	.11[c]	.03	−.10	.12[c]
	(.03)	(.04)	(.09)	(.05)
Nonrecipients of AFDC	.05	.02	.09	.04
(N = 1,692)	(.04)	(.07)	(.17)	(.07)
Partial-Year Recipients of AFDC	.21[b]	−.04	−.16	.14
(N = 206)	(.09)	(.14)	(.22)	(.14)
Full-Year Recipients of AFDC	.14[c]	−.01	−.20[a]	.18
(N = 645)	(.04)	(.06)	(.10)	(.06)
Amount of Obligation				
Full Sample (N = 2,543)	190.3	674.0[c]	−194.8	44.1
	(161.6)	(226.5)	(469.7)	(231.3)
Nonrecipients of AFDC	321.9	1,100.7[c]	−90.2	157.3
(N = 1,692)	(210.2)	(324.8)	(874.5)	(332.0)
Partial-Year Recipients of AFDC	183.5	−585.6[b]	−431.3[a]	−189.0
(N = 206)	(272.7)	(361.7)	(590.0)	(363.0)
Full-Year Recipients of AFDC	48.1	579.8[a]	−395.0	−212.0
(N = 645)	(288.0)	(345.1)	(670.6)	(375.5)

Notes: All equations contain also a set of control variables described in the text. The estimated effects are for 1981. Standard errors are shown in parentheses.

[a]Statistically significant at the 10 percent level.
[b]Statistically significant at the 5 percent level.
[c]Statistically significant at the 1 percent level.

award. Lack of a child support award is particularly prevalent among unwed mothers, who constitute one-half of the AFDC caseload. Without a formal child support award, collection is not possible.

Using the same estimation procedures as before with a slightly different variable specification, Table 5 presents the estimated effects of program services on whether the mother has a child support obligation and the amount of the obligation. The results indicate significant effects of the program on whether there is an obligation for AFDC families, but not for non-AFDC families. As before, the results seem to suggest the importance of providing the full range of services. Without following through with an attempt to establish an obligation, father locator and paternity establishment services do not appear to yield a positive payoff. With regard to dollar amounts, on the other hand, it does appear efforts to locate the father yield a positive return for nonrecipients and full-year AFDC recipients.

The lack of a significant overall effect for non-AFDC families on whether an obligation is established, and for all families on the amount of the obligation, suggests a possible mechanism for further improving the program. Recent legislation has been aimed at improving collection for families already having a formal child support award. The results presented here seem to indicate that improvements can also be made in the procedures used to establish obligations. Because a formal child support award does not exist in over one-half the families, such improvements could lead to a substantial increase in overall collections.[40] As the results

[40]Although increasing the award rate represents a potential way of increasing collections, it may not be cost effective. A rigorous determination of whether the

below indicate, such improvements are potentially as effective as those directed toward families already having an obligation.

III. Policy Implications

In the previous sections, an attempt has been made to determine the effect of current child support enforcement procedures on receipt of child support and the effect of receiving child support on welfare dependency. The results indicate that the Child Support Enforcement Program has a significantly positive effect on receipt of child support and that receipt of child support reduces welfare dependency.

Estimates of the magnitude of the effect of child support enforcement policies on welfare dependency and poverty are presented in this section by applying the results to an analysis of various types of child support systems. The objective is to compare various policy outcomes under several child support systems (including the present system). By simulating welfare dependency and poverty rates under each of these systems, an overall assessment can be made of the potential of the Child Support Enforcement Program as a mechanism for increasing the economic well being of single-parent families.[41]

In characterizing each child support system, the results from Tables 3 and 5 are used. Seven systems, each with varying degrees of program effectiveness, are examined. First, under system 1, all child support payments currently received by families are subtracted from their total income. This system portrays the worst possible situation facing single parent families; that is no help at all from the absent parent. Second, program effects from Table 3 are subtracted from all families who have reported receiving help from the child support agency. This system describes how families would fare on their own without the Child Support Enforcement Program. Third, the current system is described based on the reported data from the *CPS*.

The next four systems consider expansions of current child support policies. The first expansion applies program effects from Table 3 (those pertaining to enforcing an obligation and obtaining collection) to all families having a formal child support award but who have not yet received enforcement services from the child support agency.[42] This system represents an expansion only in collection efforts by the child support agency. Next, program effects from Table 3 are applied to all families who have not reported receiving assistance from the child support agency, including those without a formal child support award. This system, then, represents an expansion in both collection efforts and efforts to establish a formal child support award. The third expansion assumes that all support currently due is collected. This system of perfect collection represents

program is cost effective is beyond the scope of this paper.

[41] Donald Oellerich and Irwin Garfinkel (1983) also simulate the effects of various child support systems. Their analysis, however, does not consider the effect of the IV–D Program or expansions of it, nor does it incorporate behavioral responses, as is done here. Furthermore, they do not investigate the effects of various child support policies on AFDC dependency and AFDC costs. On the other hand, they consider new systems in which child support awards are increased and are tied directly to estimates of the absent parent's ability to pay. For portions of their analysis that overlap with what is presented here, the results are qualitatively similar.

[42] There are several limitations to using the estimated effects in this fashion. The most important is that IV–D effects may be different for persons who have not yet received services. If targeting of services under the current system is nonrandom, or if there are important interactions with observed characteristics of the families that are not captured in the estimates of Table 3, then these estimated effects would not be appropriate for the families not currently receiving IV–D services. Another limitation is that there may be underreporting of receipt of program services by AFDC families. A third limitation concerns applying *OLS* coefficients in the simulations, when nonlinear probit and Tobit coefficients are more appropriate. This implies that the mean effects produced by the simulations may be different from the true mean effects if those receiving services are not representative of the sample. Without additional information, it is not possible to assess the importance of these limitations on the predicted effects of this and the other systems that use the estimated program effects in Tables 3 and 5.

in some sense an upper bound of the potential of the 1984 child support amendments, assuming that the legislation does not induce families without an award to seek one. Under the final system, the behavioral equations underlying the results in Table 5 are used to predict an award amount for each family that does not currently have an award. These predicted amounts, together with the reported amounts for the other families, are then assumed to be fully collected. This last system, then, represents the maximum amount of child support obtainable within the structure of the current legal environment.

For each of the seven systems described above, three policy outcomes are examined: the AFDC participation rate, the poverty rate, and the amount of child support collected as a percentage of AFDC benefits paid. To calculate the AFDC participation rate, expected values from the AFDC participation model are calculated for each family under each system. The variables varying across each system are those involving changes in nonwage income. Results both with and without behavioral effects are reported.[43] To calculate the poverty rate, each family's total income is compared to the poverty level for that family, which is given in the *CPS*. The *CPS* defines poverty on the basis of cash income only and does not consider in-kind benefits. For purposes here, this is not an important limitation because interest is in primarily comparing poverty rates across different child support systems rather than examining the poverty rate itself. Poverty rates are not calculated for the second, fourth, and fifth systems because the method used to calculate income under these systems would lead to incorrect estimates of the poverty rate.[44]

For the poverty rate, results are presented both with and without a labor supply response. The labor supply responses are based on the income and substitution effects derived from the results of specification (3) in Table 2. The resulting labor supply responses are multiplied by the predicted wage rate to obtain an earnings response.[45]

The third policy outcome calculated is the amount of child support collected as a percentage of AFDC benefits. This is an often-quoted figure in discussions of child support policies and appears regularly in official program publications. It is intended to serve as an indicator of the collection potential of the program. This percentage is calculated in two ways; first, by dividing child support received by the family under each system to the AFDC benefit the family would receive if there were no child support, and second, by dividing child support received by the family under each system to the AFDC benefit the family would receive under that system. These calculations are only performed

[43]Specification 3 of Table 2 (which assumes a zero uncompensated wage effect) is used to generate the predictions including a behavioral effect.

[44]The reason is because each family is assigned an expected change in income (based in the results in Tables 3 and 5) and the distribution of the changes in the sample is not considered. It would be possible, using more elaborate (stochastic) simulation techniques, to estimate poverty rates under these three systems.

[45]The labor supply responses are calculated in the following manner. For nonrecipients of AFDC, child support is either taken away or added to the family's income depending on which system is being considered. Then, the income effect is applied and a new labor supply is calculated. For the system in which child support is taken away, the AFDC benefit, B, is calculated. If $B<0$ and $|B|<1/2(Wt)^2S_b$, or if $B>0$, the person is assumed to participate in the AFDC program and the appropriate substitution and income effects are applied to yield a final labor supply. Nonrecipients who were income eligible prior to the response are not assumed to join the AFDC rolls. It may be noted that the range of poverty rates may be narrower or larger than the range when labor supply response is not considered, depending on how many new AFDC participants are generated by the labor supply response. As the results below indicate, the calculated range is narrower implying that the positive labor supply responses dominate the negative labor supply responses. For AFDC recipients, no response is assumed when child support is taken away (except in South Carolina and Mississippi where AFDC taxes child support at less than 100 percent). When child support is increased, the substitution effect is applied and B is calculated at the new labor supply. If $B<0$ and $|B|>1/2(Wt)^2S_b$, then the person is assumed to leave the AFDC program and the appropriate income effect is added to the substitution effect and a final labor supply is calculated. If $B>0$ or if $|B|<1/2(Wt)^2S_b$, no response is assumed.

TABLE 6—PREDICTED EFFECTS OF CHILD SUPPORT ENFORCEMENT ON WELFARE DEPENDENCY, POVERTY, AND WELFARE COSTS

Child Support System	AFDC Participation Rate		Poverty Rate		Child Support Collections as a Percent of AFDC Benefits[a]	
	(A)	(B)	(A)	(B)	(A)	(B)
No Child Support	.35	.36	.51	.52	0 (0)	0 (0.)
Child Support—No Enforcement Program	.34	.34	–	–	–	4.8 (5.1)
Existing Child Support Enforcement Program (in 1981)	.34	.34	.48	.48	6.2 (6.7)	6.2 (6.7)
Expansion of Child Support Enforcement Program						
Full participation among those due child support	.34	.34	–	–	–	7.9 (8.6)
Full participation among all eligible families	.34	.34	–	–	–	13.5 (15.5)
Full enforcement of all existing obligations	.34	.34	.47	.47	13.8 (15.9)	13.5 (15.6)
Full enforcement of obligations for all families	.33	.32	.45	.43	32.2 (46.9)	31.6 (46.2)

Note: Columns A = With Behavioral Response; Columns B = Without Behavioral Response.
[a]Percentages measured from base with no child support. Figures in parentheses are percentages measured relative to AFDC benefits paid under the system in question.

for families who reported receiving AFDC benefits in 1981. As in the case of the other outcome measures, results are presented both with and without a behavioral response. For the results with a behavioral response, only systems 6 and 7 are considered. For these systems, the percentages are higher with a behavioral response because some families are predicted to leave the AFDC program.

The simulation results are presented in Table 6.[46] Three things stand out in this table. First, the results indicate that regardless of whether behavioral responses are allowed, AFDC dependency is quite insensitive to changes in child support policies. Moving from the worst possible situation in which no child support is collected, to the best possible situation in which the maximum amount of child support is collected for each family, reduces the AFDC dependency rate by only .04 (11 percent) when no behavioral response is allowed and by .02 (6 percent) when behavioral response is allowed. The Child Support Enforcement Program, even if it were to provide services to all eligible families is predicted to have virtually no impact on AFDC dependency.

Second, as an antipoverty device, child support enforcement again appears somewhat ineffective. Comparing the worst possible situation to the best possible situation reduces the poverty rate by .09 percentage points (17 percent) when no behavioral response is allowed, and by .06 (12 percent) when behavioral response is allowed. Within the feasible policy range, the poverty rate is only slightly affected.

Part of the reason for such a relatively small impact of child support enforcement on welfare dependency and poverty is probably low child support award amounts, particularly among AFDC recipients. In the *CPS* sample, the average child support award amount for full-year AFDC recipients in 1981 was $180 per month, or about $86 per

[46]All calculations are performed on the unweighted *CPS* sample.

child. The comparable figures for non-AFDC recipients were $207 and $127, respectively. Because the average AFDC benefit was about $282 per month and the average poverty level was about $650 per month, full enforcement of child support obligations simply will not generate enough of an increase in income (either mechanically or behaviorally) to cause many families to escape welfare dependence and poverty. Higher award amounts and/or other sources of income (principally greater earnings) are necessary.[47]

Third, while child support policies appear to have a minimal effect on welfare dependency and poverty, they do have a potentially significant effect on AFDC costs. Under the existing system, child support collections represent approximately 6.7 percent of current AFDC benefits.[48] Assuming no behavioral response, the analysis in this paper suggests that about 1.6 percent (or just under one-quarter) can be attributed directly to the IV–D Program. In the absence of the program, it is estimated that the remaining 5.1 percent would have been collected by the mothers through their own efforts.

If the child support agencies provided the full range of services to all AFDC families, it is estimated that they could recover about 15–16 percent of AFDC costs.[49] This is about the same as would be collected if there were full enforcement of all existing obligations, illustrating the potential effectiveness of greater efforts to establish obligations. If there were full enforcement of obligations for all families, close to one-half of AFDC benefits paid could be recovered. The percentage is slightly higher when behavioral response is assumed. These findings suggest that a successful system of mandatory wage withholding coupled with greater efforts to establish obligations could recover somewhat between 15 and 20 percent of AFDC benefits.[50]

IV. Conclusions

Child support enforcement has been receiving increased attention among policymakers in recent years, as evidenced by the enactment of several important pieces of legislation. One of the main purposes of such legislation is to reduce welfare costs by shifting responsibility for the support of young children from the government to the absent parent. In addition, it is hoped that child support enforcement will enable many families to eventually escape welfare dependency either by leaving the rolls or by being prevented from joining the rolls.

The analysis of this paper suggests that child support enforcement may represent an effective means for reducing AFDC program costs. However, because the current legal system establishes such low child support award amounts, it does not appear to be an effective antipoverty device. To increase the antipoverty effectiveness of child support enforcement would seem to require higher award amounts coupled with an effective mechanism for ensuring their collection. Nonetheless, even with higher award amounts, it does not appear likely that child support enforcement would have a dramatic effect on reducing the economic insecurity facing many single-parent families.

[47] Oellerich and Garfinkel find that significantly higher award amounts based on a standard applied to absent father incomes and enforcement of such awards could significantly reduce poverty.

[48] The official statistics (U.S. Department of Health and Human Services, 1983) report that child support collections were 5.2 percent of AFDC payments in fiscal year 1981.

[49] This may be somewhat of an overestimate because of the use of *OLS* coefficients to generate the responses. In other words, if those receiving services have characteristics significantly different from the mean individual in the sample, then the effect at the mean (as reflected in the *OLS* coefficients) may be greater than the true mean effect (derived from nonlinear probit and Tobit coefficients).

[50] It is important to reemphasize that these estimated reductions in AFDC costs do not take into account child support enforcement expenditures. It is possible that such expenditures would exceed collections, resulting in a net overall increase in government spending.

REFERENCES

Ashenfelter, Orley, "Determining Participation in Income-Tested Social Programs," *Journal of the American Statistical Association*, September 1983, *78*, 517–25.

Blank, Rebecca M., "The Impact of State Economic Differentials on Household Welfare and Labor Force Behavior," *Journal of Public Economics*, October 1985, *28*, 25–58.

Bumpass, Larry and Rindfuss, Ronald R., "Children's Experience of Marital Disruption," *American Journal of Sociology*, July 1979, *85*, 49–62.

Fraker, Thomas, Moffitt, Robert and Wolf, Douglas, "Effective Tax Rates in the AFDC Program," *Journal of Human Resources*, Spring 1985, *20*, 251–63.

Furstenberg, Frank F., Jr., et al., "The Life Course of Children of Divorce," *American Sociological Review*, October 1983, *48*, 656–68.

Greene, William H., "Sample Selection Bias as a Specification Error: Comment," *Econometrica*, May 1981, *49*, 795–98.

Heckman, James J., "Sample Selection Bias as a Specification Error," *Econometrica*, January 1979, *47*, 153–61.

Henderson, James M. and Quandt, Richard E., *Microeconomic Theory*, 3rd ed., New York: McGraw-Hill, 1980.

Hutchens, Robert, "Changes in AFDC Tax Rates, 1967–1971," *Journal of Human Resources*, Winter 1978, *13*, 60–74.

Moffitt, Robert, "An Economic Model of Welfare Stigma," *American Economic Review*, December 1983, *73*, 1023–35.

Moynihan, Daniel Patrick, "Children and Welfare Reform," *Journal of the Institute for Socioeconomic Studies*, Spring 1981, *6*, 1–20.

Oellerich, Donald T. and Garfinkel, Irwin, "Distributional Impacts of Existing and Alternative Child Support Systems," *Policy Studies Journal*, September 1983, *12*, 119–30.

Plant, Mark W., "An Empirical Analysis of Welfare Dependence," *American Economic Review*, September 1984, *74*, 673–84.

Robins, Philip K., "Child Support Enforcement as a Means of Reducing Welfare Dependency and Poverty," Discussion Paper No. 758–84, Institute for Research on Poverty, University of Wisconsin, October 1984.

______, "A Comparison of the Labor Supply Findings from the Four Negative Income Tax Experiments," *Journal of Human Resources*, Fall 1985, *20*, 567–82.

______ **and West, Richard W.,** "Program Participation and Labor Supply Response," *Journal of Human Resources*, Fall 1980, *15*, 499–523.

Rosett, Richard N. and Nelson, Forrest D., "Estimation of the Two-Limit Probit Regression Model," *Econometrica*, January 1975, *43*, 141–46.

Silberberg, Eugene, *The Structure of Economics: A Mathematical Analysis*, New York: McGraw-Hill, 1978.

Weiss, Yoram and Willis, Robert J., "Children as Collective Goods and Divorce Settlements," *Journal of Labor Economics*, July 1985, *3*, 268–92.

U.S. Department of Commerce, Bureau of the Census, *Child Support and Alimony: 1981 (Advance Report)*, *Current Population Reports*, Series P–23, No. 124, Washington: USGPO, May 1983.

______, *Child Support and Alimony: 1978*, *Current Population Reports*, Series P–23, No. 112, Washington: USGPO, September 1981.

______, *Statistical Abstract of the United States*, 105th ed., Washington: USGPO, 1985.

U.S. Department of Health, Education, and Welfare, Office of Child Support Enforcement, *First Annual Report to Congress on the Child Support Enforcement Program*, Washington: USGPO, June 30, 1976.

U.S. Department of Health and Human Services, (1982a) Office of Child Support Enforcement, *Child Support Enforcement: Seventh Annual Report to Congress for the Period Ending September 30, 1982*, Washington: USGPO, December 1982.

______, (1982b) Social Security Administration, Office of Research and Statistics, *Evaluation of the Child Support Enforcement Program*, Second Year Final Report prepared by Maximus, Inc., Washington:

USGPO, April 1982.

______, Office of Child Support Enforcement *Eighth Annual Report to Congress for the Period ending September 30, 1983*, Washington: USGPO, December 1983.

______, Social Security Administration, Office of Family Assistance, *Characteristics of State Plans for Aid to Families with Dependent Children*, Washington: USGPO, 1981, 1982.

U.S. House of Representatives, (1983a) Committee on Ways and Means, Subcommittee on Oversight and Subcommittee on Public Assistance and Unemployment Compensation, *Background Material on Poverty*, Washington: USGPO, October 17, 1983.

______, (1983b) Select Committee on Children, Youth, and Families, *U.S. Children and Their Families: Current Conditions and Recent Trends*, Washington: USGPO, May 1983.

On the Voluntary and Involuntary Provision of Public Goods

By B. DOUGLAS BERNHEIM*

This paper extends preexisting results concerning voluntary private funding of public goods. The assumption that individuals care about the magnitude of their own contributions only insofar as these contributions affect the aggregate level of expenditures is shown to have untenable implications. The analysis suggests that a reexamination of the factors that motivate individuals to make contributions is in order.

Many public goods, including most charitable causes and various forms of political activity, are funded predominantly through voluntary private contributions. Recently, this topic has attracted a great deal of interest.[1] In a provocative paper, Peter Warr established a simple, yet startling result: "When a single public good is provided at positive levels by private individuals, its provision is unaffected by a redistribution of income. This holds regardless of differences in individual preferences and despite differences in marginal propensities to contribute to the public good" (1983, p. 211). Warr's conclusion depends critically upon two assumptions: individuals care about the magnitude of their own contributions only insofar as these contributions affect the aggregate level of expenditures; and all individuals make strictly positive contributions. Theodore Bergstrom et al. have recently provided a more general and illuminating analysis of this proposition.

In this paper, I suggest that Warr's result as well as Bergstrom et al.'s generalizations, are only the tip of the iceberg. Maintaining Warr's assumptions, I demonstrate the following:

1) Any policy consisting of lump sum transfers, and lump sum public contributions to the privately provided public good, has no effect on resource allocation.

2) Any policy consisting of apparently "distortionary" transfers, and distortionary public finance of the privately provided public good, has no effect on resource allocation.

3) In general, the method used to raise revenue for public transfers and expenditures (including expenditures on public goods for which *no* private contributions are made) is completely irrelevant: all taxes are equivalent to lump sum taxes.

One might be tempted to dismiss these results on the grounds that, while the vast majority of individuals contribute to some cause, no single privately provided public good receives universal support. However, it turns out that this observation is immaterial, as long as there is sufficient overlap between the sets of donors to different causes (for example, A contributes to the Republicans, B to the Democrats, C to the Republicans and the United Way, and D to the Democrats and the United Way). Furthermore, Kyle Bagwell and I (1985) have argued that voluntary interpersonal transfers serve to link different individuals in the same manner as voluntary contributions to public good.[2]

*Department of Economics, Stanford University, Stanford, CA 94305. I thank Lawrence Summers for helpful discussions.

[1]Theodore Bergstrom, Lawrence Blume, and Hal Varian (1984) provide a brief but comprehensive review of this literature.

[2]In my paper with Bagwell, we focus attention on transfers between individuals, and do not explicitly consider the role of public goods. Of course, one can think about a public project as a selfish individual who cares only about his own consumption, and who therefore makes transfers to no one else. To the extent a number of individuals contribute voluntarily to the project, our analysis is applicable (thus, in our textual example, we may think of the Democrats as an individual, to whom both B and D contribute). However, one

Thus, it seems quite plausible that everyone who either makes or receives a voluntary transfer or contribution is indirectly linked through such transactions to everyone else, in which case the conditions for widespread private neutralization of public policies are satisfied.

What, then, are we to make of these results? Rather than believe that virtually all government policy is neutral, I am inclined to reexamine the critical assumptions. Two alternatives are available. First, one could argue that large segments of the population are isolated, in the sense that they are not linked through chains of voluntary transfers and contributions. While this view may seem inviting, it is troubling in one respect. Suppose, hypothetically, that all individuals were linked through such chains. Do we indeed believe that this would neutralize the effects of all taxes and transfers? If one's answer is negative, then it is necessary to examine other assumptions.

As mentioned earlier, these results also depend critically upon the assumption that individuals care about the magnitude of their own contributions only insofar as these contributions affect the aggregate level of expenditures. This explicitly rules out the possibility that individuals enjoy (or dislike) making transfers, or that benefactors acquire bargaining leverage over recipients. Elsewhere, Andrei Shleifer, Lawrence Summers, and I (1985) have argued that there is strong empirical support for such alternative formulations. In order to avoid the implausible implications of equilibrium behavior in models of contributions and transfers, one is naturally and inevitably led to these alternatives.

I. The Model

Consider an economy consisting of i individuals, indexed $i=1,\dots,N$. Each selects a level of labor supply (l_i), consumes a private good (χ_i), and enjoys the benefits of two public goods (G, H). I write i's utility as a general function of these variables:

$$u_i = u_i(l_i, \chi_i, G, H). \tag{1}$$

The units of l_i, G, and H were chosen so that all prevailing prices are unity.[3]

The allocation of resources is determined in three stages. In stage 1, the government picks a policy,

$$P = (y, \tau, \gamma_0, \eta_0). \tag{2}$$

The policy consists of four distinct components. First, the government selects a vector of exogenous (nonlabor) incomes for the consumers,

$$y \equiv (y_1, \dots, y_N). \tag{3}$$

Implicitly, in choosing y, the government determines lump sum taxes and transfers. I denote the total amount of nonlabor income available in this economy as Y. Second, the government selects a vector of income tax functions,

$$\tau = (\tau_1, \dots, \tau_N). \tag{4}$$

I allow τ_i to depend on the entire vector of labor supply decisions, $l \equiv (l_1, \dots, l_N)$: $\tau_i(l)$ indicates i's income tax payment when individuals have chosen to supply l. Among other things, this formulation permits the government to condition its transfer policies upon available revenues. Third, the government selects a function, γ_0, which describes its contributions to the first public good as a function of the labor supply vector (national income). Similarly, η_0 is a function which describes contingent contributions to the second public good.

The government must, of course, respect its budget constraint under all contingencies.

requires a more fully articulated model, such as that considered here, to analyze the effects of financing public goods which are supported exclusively by the government.

[3] In this respect, my analysis does not treat a general equilibrium. However, the extension to general equilibrium is straightforward: to the extent policies are allocatively neutral, any prices which prevailed prior to the policy may also prevail after.

Specifically, for all l,

$$(5)\quad \eta_0(l)+\gamma_0(l)=\left(Y-\sum_{i=1}^{N} y_i\right)+\sum_{i=1}^{N}\tau_i(l).$$

In stage 2, each consumer selects his labor supply, and realizes his total net income:

$$(6)\qquad y_i+l_i-\tau_i(l).$$

At the same time, government collects its revenue.

In stage 3, the government funds the public goods as prescribed by its policy P. Consumers divide their resources between the private good, χ_i, and contributions to the public good. In order to illustrate basic principles within as simple a context as possible, I assume that all individuals contribute to the first public good (G), and that none contribute to the second (H). This configuration of contributions may arise due to the structure of preferences for G and H, or may be the consequence of government policy (for example, government crowds out private contributions to H, but provides no funding for G). For concreteness, we may think of G as funding of religious organizations, while H represents provision for the national defense.

I will use g_i to denote i's contributions to the first public good. Ordinarily, one would introduce nonnegativity constraints, $g_i \geq 0$, as in Bergstrom et al. However, I wish to assume that these constraints are nonbinding both before and after the imposition of some particular policy. In such cases, nonnegativity constraints do not affect the local opportunities of any individual. While they do affect global opportunities, it is permissible to ignore them as long as either 1) the economy has sufficient convexity properties, or 2) the economy has sufficient continuity properties, and the policy represents a small change from the prevailing system. I will assume that one of these two conditions is satisfied, and henceforth consider a world in which contributions may be negative.

The allocation procedure described above induces a game between consumers. Given the policy P^*, consumer i's strategy consists of announcing a level of labor supply, l_i, and a function, γ_i, which prescribes a level of contributions for i in stage 3 for every potential labor supply vector, l, chosen in stage 2. Let $\gamma \equiv (\gamma_1,\ldots,\gamma_N)$. Let us say that (l^*,γ^*) is an *equilibrium*[4] if, for all l, $\gamma_i^*(l)$ solves

$$(7)\quad \max_{g_i} u_i\left(l_i,\, y_i^*+l_i-\tau_i^*(l)-g_i,\; \sum_{j\neq i}\gamma_j^*(l)+g_i,\, \eta_0^*(l)\right),$$

where "$j \neq i$" is understood to include the case of $j=0$, and if l_i^* solves

$$(8)\quad \max_{l_i} u_i\left(l_i,\, y_i^*+l_i-\tau_i^*(l_i,l_{-i}^*) - \gamma_i^*(l_i,l_{-i}^*),\; \sum_{j=0}^{N}\gamma_j^*(l_i,l_{-i}^*),\; \eta_0^*(l_i,l_{-i}^*)\right).$$

A final word concerning the model. To obtain the general result described below, it is essential that contributions to the public good be chosen *subsequent* to the choice of labor supply. While this may appear restrictive, it is actually relatively innocuous. In a more general context, my result would require only that individuals are linked through *some* network of transfers and contributions subsequent to each potentially distorted choice.

II. The Result

I am now prepared to prove the central result.

THEOREM 1: *Consider any two policies,* $P=(y,\tau,\gamma_0,\eta_0)$, *and* $P'=(y',\tau',\gamma_0',\eta_0')$. *Suppose* $\eta_0=\eta_0'$. *Then any final allocation sustained as an equilibrium under P can also be sustained as an equilibrium under P'.*

This result indicates that, as long as the government adheres to its policy for allocat-

[4] Since we require that $\gamma_i(l)$ solves (7) for all l, this definition corresponds to Reinhard Selten's (1965, 1975) notion of subgame perfect Nash equilibrium (each l defines a distinct subgame).

ing the second (publically provided) public good (η_0), altering lump sum taxes and transfers (y), distortionary taxes (τ), and/or contributions to the first (privately provided) public good (γ_0) will have no effect on the private consumption of any individual, nor on chosen labor supplies, nor on the total levels of public goods provided.

PROOF:

Suppose that, under P, we have some equilibrium (l^*, γ). Now consider P'. Define the vector of new strategies

$$(9)\quad \gamma_i'(l) = \gamma_i(l) + (y_i' - y_i) - (\tau_i'(l) - \tau_i(l)).$$

I will show that (l^*, γ') is an equilibrium under P'. It is easy to verify that it induces the same final allocation (private goods and public goods) as does (l^*, γ) under P.

Take some l. Suppose $j \neq i$ uses γ_j'. Then i's problem is to

$$(10)\quad \max_{g_i'} u_i\left(l_i, y_i' + l_i - \tau_i'(l) - g_i', \sum_{j \neq i} \gamma_j'(l) + g_i', \eta_0(l)\right).$$

Now do the following change of variables:

$$(11)\quad z = g_i' - (y_i' - y_i) + (\tau_i'(l) - \tau_i(l)).$$

Individual i's problem is then to

$$(12)\quad \max_{z} u_i\Bigg(l_i, y_i + l_i - \tau_i(l) - z, \quad \gamma_0'(l) + \sum_{j \neq 0, i} \gamma_j(l) + \sum_{j=1}^{N} [(y_j' - y_j) - (\tau_j'(l) - \tau_j(l))] + z, \eta_0(l)\Bigg)$$

Using (5) for both P and P', we see that

$$(13)\quad \sum_{j=1}^{N} [(y_j' - y_j) - (\tau_j'(l) - \tau_j(l))] = \gamma_0(l) - \gamma_0'(l).$$

Substituting this into (12), we see that i's problem is to

$$(14)\quad \max_{z} u_i\left(l_i, y_i + l_i - \tau_i(l) - z, \sum_{j \neq i} \gamma_j(l) + z, \eta_0(l)\right).$$

But since (l^*, γ) is an equilibrium under P, we know that $z = \gamma_i(l)$ is a solution to (14). Thus, returning to the original variables, it can be seen that $g_i' = \gamma_i'(l)$ solves (10). Thus, condition (7) is satisfied for all l.

Now consider condition (8). I wish to show that l^* solves

$$(15)\quad \max_{l_i} u_i\left(l_i, y_i' + l_i - \tau_i'(l_i, l_{-i}^*) - \gamma_i'(l_i, l_{-i}^*), \sum_{j=0}^{N} \gamma_j'(l_i, l_{-i}^*), \eta_0(l_i, l_i^*)\right).$$

Using (9) and (13), it is easy to see that

$$(16)\quad \sum_{j=0}^{N} \gamma_j'(l_i, l_{-i}^*) = \sum_{j=0}^{N} \gamma_j(l_i, l_{-i}^*).$$

Substituting (16) and (9) into (15), it can be seen that the consumers' problem is to

$$(17)\quad \max_{l_i} u_i\left(l_i, y_i + l_i - \tau_i(l_i, l_{-i}^*) - \gamma_i(l_i, l_{-i}^*), \sum_{i=0}^{N} \gamma_j(l_i, l_{-i}^*), \eta_0(l_i, l_{-i}^*)\right).$$

But, since (l^*, γ) is an equilibrium under P, l_i^* solves (17). Consequently, conditions (7) and (8) are both satisfied—(l^*, γ') is an equilibrium under P'.

Intuitively, Theorem 1 holds for a remarkably simple reason: if all but one individual acts to offset the policy change for each labor supply profile, then the opportunity set of the remaining individuals (achievable combinations of l_i, χ_i, G, and H) are unchanged. It is therefore optimal for him, as well, to neutralize any effects.

It is natural to wonder whether this result is sensitive to the introduction of several privately provided public goods. In particular, while each individual might contribute to some good, it is easily conceivable that no single good would receive universal support. Through a completely analogous argument, one can show that the central result continues to hold as long as it is impossible to partition the set of consumers into two subsets, I_1 and I_2, and the sets of goods into two subsets, G_1 and G_2, such that members of I_i contribute *only* to goods in G_i $(i=1,2)$. This condition implies that one can find a chain of contributions linking any two members of the population.

To reiterate, the result depends essentially upon only two assumptions. First, individuals care about the magnitude of their contributions only insofar as these contributions affect the aggregate level of expenditures. Second, chains of operative voluntary transfers and contributions link all individuals (my paper with Bagwell extends the argument provided here to situations in which linkages are indirect). For reasons described earlier, the first of these assumptions deserves much closer scrutiny.

REFERENCES

Bergstrom, Theodore C., Blume, Lawrence and Varian, Hal R., "On the Private Provision of Public Goods," Working Paper, University of Michigan, December 1984.

Bernheim, B. Douglas and Bagwell, Kyle, "Is Everything Neutral?," mimeo., Stanford University, 1985.

______, Shleifer, Andrei and Summers, Lawrence, "The Strategic Bequest Motive," *Journal of Political Economy*, December 1985, *93*, 1045–76.

Selten, Reinhard, "Spieltheoretische Behandlung eines Oligopolmodells mit Nashfragetragheit," *Zeitschrift fur Gesamte Staatswissenschaft*, 1965, *12*, 301–24.

______, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, January 1975, *4*, 25–55.

Warr, Peter G., "The Private Provision of a Public Good is Independent of the Distribution of Income," *Economic Letters*, 1983, *13*, 207–11.

Educational Achievement In Segregated School Systems: The Effects of "Separate-but-Equal"

By ROBERT A. MARGO*

In 1896, the Supreme Court ruled that racially "separate-but-equal" public facilities, including schools, were constitutional. In 1954, the Court reversed the decision. The willful and flagrant violations of the separate-but-equal doctrine in southern public schools have been documented extensively.[1] Evidence from the segregation era linking educational discrimination to educational outcomes, however, has been examined far less. Using data from 1920 to 1940, I show that closing the racial gap in school inputs would have raised the educational achievement of black children relative to white children. But low black incomes, wealth, and high rates of adult illiteracy helped sustain a significant racial achievement gap even if separate-but-equal were reality instead of myth.

I. Empirical Analysis

My model of educational achievement is the conventional one (Eric Hanushek, 1972; Anita Summers and Barbara Wolfe, 1977):

$$A = f(M, X, E), \tag{1}$$

where A indicates achievement, M measures student motivation, X is a vector of school inputs, and E is a vector of socioeconomic variables. A version of equation (1) is estimated using race-specific county averages for Alabama in 1920, 1930, and 1940 (see Table 1 for data sources).

Achievement is captured by the literacy rate of children and young adults, aged 7 to 20.[2] More than any other type of human capital, acquisition of literacy by blacks profoundly influenced the historical evolution of racial income and occupational differences (Robert Higgs, 1977; James Smith, 1984). The Alabama literacy rates are not specific to public school students, but the bias is probably small.[3]

Student motivation is proxied by the average daily attendance rate of pupils enrolled in grades one through six.[4] School inputs are: the length of the school year in days; per pupil expenditures on teacher salaries, per school day, a measure of instructional quality; the per pupil value of school capital; and the proportion of one-teacher schools. Finis Welch argued that "discipline would have consumed a significant proportion of instructional time and energy" (1973, p. 59) in one-teacher schools, so the coefficient of this variable should be negative.

Socioeconomic inputs are race (equation (1) is estimated separately by race), and

*Department of Economics, Colgate University, Hamilton, NY 13346, and National Bureau of Economic Research. This paper was written while I was a member of the Department of Economics, University of Pennsylvania. Research support from the University of Pennsylvania is gratefully acknowledged. I am grateful to Orley Ashenfelter, Stanley Engerman, Claudia Goldin, Anita Summers, Paul Taubman, seminar participants at the University of Pennsylvania and Colgate University, and two anonymous referees for helpful comments. Any errors are my own. This research is part of the NBER's Development of the American Economy program. Any opinions expressed are my own and not those of the NBER. An earlier version of this paper (1985b) appeared as an NBER working paper.

[1]See, for example, Horace Mann Bond (1934, 1939), Louis Harlan (1958), Richard Freeman (1972), Finis Welch (1973), J. Morgan Kousser (1980), and my 1985a book.

[2]Many southern states reported race-specific information on school inputs, but to the best of my knowledge, only Alabama reported race-specific, county-level literacy rates for the period.

[3]Private school enrollments for both races were no more than 5 percent of public school enrollments over the period. See State of Alabama (1920, p. 50; 1940, pp. 386–88).

[4]A better proxy would be the average daily attendance rate of the elementary school-age population, but consistent county-level data are not available for all three census years.

proxies for income and wealth. They are included because of their importance in modern studies of achievement (Hanushek; Summers and Wolfe), and because of the large racial economic differences that prevailed in the early twentieth-century South. Race-specific, county-level data on incomes are not available for the period, and must be estimated. The estimation procedure applies fixed race-specific income weights derived from the 1950 Census to urban, rural farm, and rural nonfarm population shares.[5] Regressions are also reported which include the urban and rural nonfarm population shares instead of the income variable. Because urban income was higher than rural nonfarm income (which was higher than rural farm income), an increase in percent urban should have a larger positive effect on literacy than an increase in percent rural nonfarm. Wealth is proxied by the race-specific share of families who own their homes.[6]

The limitations of the data are numerous. Their aggregate nature obscures the heterogeneity of inputs and outputs within counties. School quality may be poorly measured by the current value of school inputs if educational change is rapid, or intercounty migration of school-age children is high. Omitted socioeconomic factors may bias the school input coefficients upwards if the left-out variables are positively correlated with literacy and school quality. Without detailed pupil-specific data and a much richer list of school and socioeconomic inputs, the biases cannot be investigated, except indirectly (see below).

Table 1 exhibits race-specific arithmetic sample means. Judging by the levels of school inputs, Alabama's white schools were much better than her black schools. The white literacy rate exceeded the black literacy rate by 20 percentage points in 1920, but the gap had fallen to 7 percentage points in 1940. Perhaps the most important change in the black schools over time was the 55 percent increase in the length of the school year. Assuming an attendance rate equal to the average, a black pupil age 12 in school continuously since age 6 would have 2.6 additional years of schooling in 1940 than in 1920.

For estimation purposes the data were pooled across census years. The dependent variable is the log of the literacy rate.[7] Except for variables in percent form, all independent variables are measured in logs. Table 2 presents random and fixed-effects estimates of the regression coefficients. The random effects disturbance term consists of a county-specific error v_j with zero mean variance σ_j^2; a year-specific error w_t with zero mean and variance σ_t^2; and an independently and identically distributed error e_{jt} with zero mean and variance 0_e^2.[8] The covariance of the random effect error and the independent variables is assumed to be zero. The fixed-effects estimator corresponds to a regression with county and census year

[5]Let s_{rf}, s_{rnf}, and s_u be, respectively, the rural farm (rf), rural nonfarm (rnf), and urban (u) proportions of the population of county j in year t. Per capita income is estimated as $y = y_{rf}s_{rf} + y_{rnf}s_{rnf} + y_u s_u$ where the y's are a series of relative income weights derived from U.S. Department of Commerce (1952). For whites, the weights are $y_{rf} = 1.00$, $y_{rnf} = 1.58$, and $y_u = 3.89$. For blacks, the weights are $y_{rf} = 0.350$, $y_{rnf} = 0.671$, and $y_u = 1.65$. The derivation of the weights is available from the author on request. The rural farm, rural nonfarm, and urban proportions for 1930 and 1940 were calculated from information in U.S. Department of Commerce (1932, 1943). Urban-rural proportions for 1920 were estimated from information in U.S. Department of Commerce (1921, 1932); details of the estimation procedure are available from the author on request. Because of errors inherent in the calculation (variations in y only reflect variations in urban-rural population shares), the income coefficients in Table 2 are biased towards zero.

[6]Race-specific, county-level data on homeownership are available for 1910, 1930, and 1940 in U.S. Department of Commerce (1918, 1935, 1943). The 1920 figures for blacks were interpolated from 1910 and 1930 data. This procedure yielded implausibly high estimates of white homeownership in 1920, and an alternate procedure—extrapolating backward at the rate of growth in white homeownership between 1930 and 1940—was used instead. Measurement error in the white homeownership figures may be responsible for the insignificant white coefficients (see the text and Table 2).

[7]The substantive conclusions are not affected if the logistic transformation of the literacy rate is used instead as the dependent variable.

[8]See Wayne Fuller and George Battese (1974) for further discussion of the random effects model and details of the estimation procedure.

TABLE 1—ARTITHMETIC SAMPLE MEANS

Variable	White 1920	White 1930	White 1940	Black 1920	Black 1930	Black 1940
Literacy Rate, Ages 7–20	0.88	0.93	0.96	0.68	0.77	0.88
Length of School Year (days)	130	151	148	93	119	141
Instructional Expenditures, per Pupil, per Day (1930$)	0.07	0.12	0.17	0.02	0.05	0.08
Value of School Capital per Pupil (1930$) $\times 10^{-2}$	0.23	1.07	1.43	0.08	0.21	0.25
Percent One-Teacher Schools	0.53	0.32	0.20	0.84	0.61	0.53
Per Capita Income	1.48	1.62	1.67	0.57	0.61	0.68
Percent Urban	0.12	0.15	0.16	0.11	0.14	0.20
Percent Rural Nonfarm	0.24	0.25	0.25	0.24	0.23	0.24
Percent Own Home	0.43	0.43	0.42	0.22	0.23	0.24

Source: All school data are from State Superintendent of Public Instruction, State of Alabama (1920, 1930, 1940). Percent urban, percent rural nonfarm, percent own home: calculated from information in U.S. Department of Commerce (1918, 1921, 1932, 1943).
Notes: Teacher salaries and school capital are deflated to 1930 dollars using the Warren-Pearson wholesale price index (U.S. Department of Commerce, 1975). Fn. 5 describes the construction of the per capita income variables.

TABLE 2—LITERACY REGRESSIONS: ALABAMA, 1920 TO 1940

Variable	White *RE*	White *FE*	White *RE*	White *FE*	Black *RE*	Black *FE*	Black *RE*	Black *FE*
Constant	−0.53	−0.43	−0.51	−0.40	−1.29	−1.23	−1.34	−1.22
	(4.10)	(2.62)	(3.87)	(2.45)	(6.17)	(5.03)	(6.62)	(5.20)
Attendance Rate	0.02	0.01	0.02	0.003	0.15	0.13	0.15	0.12
	(1.03)	(0.21)	(1.01)	(0.12)	(2.93)	(2.09)	(2.88)	(1.95)
Length of School	0.10	0.07	0.10	0.07	0.25	0.18	0.25	0.17
Year (in days)	(4.25)	(2.40)	(3.89)	(2.39)	(6.62)	(3.54)	(6.55)	(3.42)
Expenditures per Pupil,	0.33	0.17	0.33	0.20	0.48	0.12	0.49	0.12
per Day $\times 10^{-1}$	(3.52)	(1.35)	(3.47)	(1.58)	(2.32)	(0.40)	(2.29)	(0.38)
Value of School Capital	0.008	−0.03	0.01	−0.02	−0.04	−0.16	−0.04	−0.16
per Pupil $\times 10^{-1}$	(0.17)	(0.55)	(0.22)	(0.41)	(0.38)	(1.53)	(0.44)	(1.52)
Percent One-	0.003	0.01	0.001	0.01	−0.03	0.03	−0.03	0.03
Teacher Schools	(0.16)	(0.68)	(0.06)	(0.69)	(0.74)	(0.52)	(0.75)	(0.56)
Per Capita Income	0.02	0.04			0.05	0.03		
	(1.98)	(1.32)			(2.10)	(0.40)		
Percent Urban			0.04	0.08			0.10	−0.008
			(2.06)	(1.21)			(2.10)	(0.06)
Percent Rural Nonfarm			0.03	−0.02			0.01	−0.13
			(1.20)	(0.33)			(0.24)	(0.70)
Percent Own Home	0.03	−0.06	0.03	−0.07	0.30	0.26	0.31	0.28
	(1.02)	(0.95)	(0.86)	(1.03)	(3.05)	(0.76)	(2.99)	(0.81)
N	201	201	201	201	180	180	180	180
MSE	0.0072	0.0077	0.0071	0.0072	0.0063	0.0063	0.0063	0.0063
Hausmann Statistic	15.1		10.8		14.1		10.8	
Significance Level	0.03		0.22		0.05		0.21	

Source: See text and Table 1.
Notes: Absolute value of *t*-statistics are shown in parentheses. Dependent variable is log of child literacy rate, ages 7 to 20. Except for variables in percent form, all independent variables are measured in logs; *RE*: random effects; *FE*: fixed effects; *MSE*: means squared error. Significance level refers to Hausmann test statistic.

dummy variables. Judging by the mean squared error (*MSE*) statistics, the regressions fit the data well, in view of the small sample size and data quality. Broadly speaking, the results affirm the model of educational production described by equation (1). I focus first on the random effects estimates.

By far the most important school inputs were the length of the school year and instructional expenditures. Both variables were significant determinants of literacy rates regardless of race. By contrast, school capital and percent one-teacher schools had no significant impact on literacy.[9] Average daily attendance rates were positively associated with literacy rates, but the effect was statistically significant only for blacks.

Race and economic factors were also important determinants of literacy rates. The income coefficients were positive and significant for both races. When percent urban and percent rural nonfarm are substituted for the income variable, the coefficient of percent urban exceeds the coefficient of percent rural nonfarm in every equation, consistent with the difference between urban and rural incomes. Homeownership had a significant positive effect on black literacy, and a positive, but insignificant, effect on white literacy. The large positive difference between the white and black constant terms suggests that Alabama's white schools were more efficient overall in transforming educational inputs into achievement.[10]

If the county-specific error were correlated with the independent variables, for example, because of differences across counties in tastes for schooling or in unobserved school or socioeconomic inputs, the random effects coefficients will be biased and inconsistent, and the fixed effects estimator is preferable on statistical grounds (Jerry Hausmann, 1978). Except for the length of the school year and the black attendance rate, none of the fixed effects coefficients is statistically significant at conventional levels. Furthermore, the fixed effects coefficients of the length of the school year and instructional expenditures are smaller in absolute value than the random effects coefficients. Judging by a Hausmann test, specification error is present in the random effects equations (see Table 2).[11] Analysis in Sections II and III makes use of both estimators, but in light of the specification test statistics, results based on the random effects coefficients should be interpreted cautiously.

II. The Effects of Separate-but-Equal

What was the effect of school inputs on the racial literacy gap in Alabama? Answering this question requires a definition of separate-but-equal. An historically defensible position is "equal school inputs," that is, reduce the racial gap in school inputs to zero.[12] The percentage of the racial literacy gap accounted for by race differences in school inputs simulates the counterfactual impact of separate-but-equal. These per-

[9]School capital was generally valued at historical cost. County figures also typically exclude privately owned buildings used as schools, which were common in black school districts. The insignificant school capital coefficients may be due to measurement error rather than absence of an underlying relationship. The negative effects of one-teacher schools may have been offset by a positive impact on younger children by mixing them with children in older grades.

[10]The racial difference in constant terms might also reflect a larger share of white children acquiring literacy before beginning school. Data for 1940 show, however, that white and black children had equal literacy rates prior to entering the first grade. See State of Alabama (1940, p. 22).

[11]The omission of the proportion of school-age children in school is probably the major source of specification error. Between 1920 and 1940, the enrollment rate of black children in Alabama ages 5 to 14 increased by 17 percentage points; the corresponding increase for white children was 7 percentage points. Changes in compulsory schooling laws may have played an independent role in eradicating child illiteracy by raising school attendance. After 1915, Alabama school code required "every child between the ages of 7 and 16 to attend school the entire length of term" (Alabama Educational Survey Commission, 1945, p. 105). Enforcement of the statute, however, was evidently lax. In 1930, expenditures on enforcement of attendance in white schools was 1.8 percent of total white instructional expenditures; the corresponding figure for blacks was 1.4 percent (State of Alabama, 1930, pp. 234, 274). As late as 1945, teachers cited nonenforcement of attendance laws as a major weakness of Alabama schools (Alabama Educational Survey Commission, p. 62).

[12]This seems to be the interpretation favored by state courts; see Charles Magnum (1940, pp. 88–90).

TABLE 3—EFFECTS OF SEPARATE-BUT-EQUAL
(Shown in percent)

Variable	1920 White	1920 Black	1930 White	1930 Black	1940 White	1940 Black
Length of School Year:						
Random Effects	12.3	30.7	12.5	31.3	5.2	13.0
Fixed Effects	8.6	22.1	8.8	22.6	3.6	9.4
Instructional Expenditures:						
Random Effects	11.7	17.0	14.1	20.5	28.6	41.6
Fixed Effects	6.0	4.2	7.3	5.1	14.7	10.4
Total School Inputs:						
Random Effects	24.0	47.7	26.6	51.8	33.8	54.6
Fixed Effects	14.6	26.3	16.1	27.7	18.3	19.8
Per Capita Income and Homeownership:						
Random Effects	9.3	40.6	13.2	57.3	25.1	106.4
Fixed Effects	9.4	30.5	13.8	42.5	27.0	79.3
Racial Literacy Gap	0.273		0.190		0.093	

Notes: Figures give percentages of racial literacy gap accounted for by racial differences in educational inputs. The formula is $B_1(X_w - X_b)/(L_w - L_b)$ where i = white, black, B_1 is the regression coefficient (equations 1,3,5,7 in Table 2), X is the educational input, and L is the literacy rate. X and L are measured in logs. All calculations are performed at the arithmetic sample means (see Table 1).

centages are shown in Table 3 for the length of the school year and instructional expenditures. The columns labelled Black are based on the black coefficients and those labelled White on the white coefficients.

Although the percentages vary depending on the year and coefficients used, an average of 27 percent of the racial literacy gap is explained by race differences in school inputs. The largest impact is indicated by the black random effects coefficients (48 to 55 percent) and the smallest by the white fixed-effects coefficients (15 to 18 percent). Enforcement of separate-but-equal would have improved the relative educational achievement of black children. But a significant literacy gap would have remained, and the row labelled Per Capita Income and Homeownership shows that a large chunk of the residual is attributable to racial economic differences. Then, as now, poverty influenced achievement in many ways, for example, by reducing the frequency of school attendance (see my 1986 paper) and the availability of reading materials in the home. An important causal factor in black poverty was adult illiteracy, which may have influenced child literacy directly. Illiterate parents could never substitute for inadequate schools and teach their children to read and write. The diffusion of literacy among blacks was essentially a cohort phenomenon. Adult black illiteracy—a legacy of slavery and educational backwardness in the late nineteenth and early twentieth-century South—remained high, absolutely and relative to white levels, until World War II (Smith).

Unfortunately, the Alabama data cannot reveal the relationship between adult and child literacy, but state-level data can. The following regression for 1930 relates the log of the black literacy rate, ages 10 to 14 ($LLIT$) to an estimate of years of schooling from age 6 to 9 (YRS), per pupil expenditures on school inputs (EXP), the per pupil value of school capital (VCP), and the literacy rate of adults, ages 35 and up ($ALIT$).[13] The sample consists of 16 southern states with legally segregated school systems and the District of Columbia. A "hat"

[13] YRS is adjusted for across-state differences in the length of the school year. The data are from U.S. Department of Commerce (1932) and Tuskegee Institute (1930).

indicates the natural logarithm of a variable (absolute values of t-statistics are shown in parentheses).

$$(2)\quad LLIT = \underset{(2.50)}{-0.10} + \underset{(1.65)}{0.04}\widehat{YRS} + \underset{(1.95)}{0.01}\widehat{EXP} - \underset{(0.55)}{0.002}\widehat{VCP} + \underset{(1.75)}{0.12}ALIT$$

$$N = 17,\ R^2 = 0.87.$$

Holding constant school inputs and years of schooling, adult literacy had a positive impact on child literacy, significant at the 10 percent level. A similar regression for whites yielded an adult literacy coefficient of 0.15 ($t = 2.67$). Equation (2) can also be used to assess the effects of separate-but-equal. Eliminating state-level race differences in school inputs accounts for 31 percent of the racial literacy gap, roughly the same as in Alabama in 1930 (see Table 3).[14]

III. The Equal Valuation Hypothesis

That southern school officials allocated fewer resources to black schools is irrefutable. There are two ways to explain their actions. Either they valued black and white achievement equally but the productivity of school inputs was lower in black schools, or they valued black achievement less than white achievement. If the "equal" valuation hypothesis were true, school officials would equate the black-to-white ratio of marginal products of schools inputs to the black-to-white ratio of school input prices. If it were false, the black-to-white ratio of marginal products would exceed the black-to-white input ratio.[15] The Alabama production functions can distinguish between the two alternatives.

Table 4 gives the black-to-white ratio of marginal products for the length of the school year.[16] To a first approximation, the per pupil marginal cost of lengthening the school year is instructional expenditures per day. Black-to-white input prices ratios based on this approximation are shown in Table 4. Dividing the input price ratios by the ratios of marginal products estimates the relative (marginal) valuation of black literacy.[17] The figures in parentheses give lower and upper bounds to the relative valuation based on 95

[14]School inputs include: expenditures per pupil, the value of the school capital stock, and the length of the school year. Details of the calculation are available from the author on request.

[15]Let n_i and A_i be, respectively, race-specific enrollment and achievement per pupil, i = white, black. For simplicity, assume that $A_i = A_i(s_i)$, where s_i is a single school input. The price of s_i is p_i. The school board budget Z is exogenously determined. A school board that values black and white achievement equally on average and at the margin is a school board that maximizes aggregated achievement subject to its budget constraint:

$$\max_{s_w, s_b} n_w A_w(s_w) + n_b A_b(s_b)$$

subject to $\quad n_w p_w s_w + n_b p_b s_b = Z.$

The first-order condition is $MP_b/MP_w = p_b/p_w$, where MP_i is the marginal product of the school input. Note that, at the optimum, $s_b \neq s_w$ unless the production functions and input prices were the same. If the board valued black and white achievement differently, the first-order condition would be

$$(MV_b/MV_w) \times (MP_b/MV_w) = p_b/p_w,$$

where MV_i is the (marginal) value of achievement. If $MV_b < MV_w$, then MP_b/MP_w will exceed p_b/p_w.

[16]The formula for the marginal products is $MP = BL/X$, where B is the regression coefficient (see Table 2), L is the literacy rate, and X is the school input. The formula assumes school boards valued the actual literacy rate and not its logarithm. All calculations are made at the arithmetic sample means (see Table 1). Substantive results are not affected if instructional expenditures per pupil is used instead of the length of the school year.

[17]Recent studies (Richard Butler, 1983; my 1984 paper) suggest a portion of lower instructional expenditures in black schools represented wage discrimination against black teachers; if so, my calculations of the relative valuation of black literacy would be biased downward. The black-to-white ratio of marginal products, however, exceeds unity in every year (see Table 4). Thus even if the black-to-white ratio of input prices were one (i.e., all of the racial difference in teacher salaries reflected wage discrimination against black teachers, and race differences in class size are ignored), Alabama school boards still valued black literacy less at the margin than white literacy.

TABLE 4—RELATIVE VALUATION OF BLACK LITERACY

Variable	1920	1930	1940
Length of School Year, Black-to-White Ratio of Marginal Products:			
Random Effects	2.70	2.63	2.41
Fixed Effects	2.78	2.70	2.47
Length of School Year, Black-to-White Ratio of Input Prices	0.38	0.44	0.45
Relative Valuation of Black Literacy × 100 Percent[a]			
Random Effects	14.1	16.1	18.7
	(5.8, 29.0)	(7.0, 34.9)	(7.8, 39.1)
Fixed Effects	13.7	16.3	18.2
	(1.6, 55.9)	(1.9, 66.7)	(2.2, 73.8)

[a]Shown in percent. The figures in parentheses give lower and upper bounds to relative valuation assuming a 95 percent confidence interval around the input coefficients.

percent confidence intervals around the school input coefficients.

The evidence in Table 4 refutes the "equal valuation" hypothesis for Alabama. The black-to-white ratio of marginal products substantially exceeds the input price ratios in all three years. Despite increases over time, Alabama school officials judged black literacy to be worth far less (at the margin) than white literacy. Clearly, dramatic changes in racial attitudes were necessary before the officials would have voluntarily complied with separate-but-equal.

IV. Concluding Remarks

This paper has analyzed the relationship between educational inputs and outputs in legally segregated school systems. Evidence for Alabama from 1920 to 1940 reveals that strict enforcement of the Supreme Court's 1896 separate-but-equal ruling, particularly with respect to the length of the school year, would have narrowed the literacy gap between white and black children. But separate-but-equal was not enough. Only a radical redistribution of school board budgets would have compensated for the poverty and adult illiteracy that hindered black school achievement in the early twentieth-century South.

REFERENCES

Bond, Horace Mann, *The Education of the Negro in the American Social Order*, New York: Prentice-Hall, 1934.

______, *Negro Education in Alabama: A Study in Cotton and Steel*, Washington: Associated Publishers, 1939.

Butler, Richard J., "Direct Estimates of the Demand for Race and Sex Discrimination," *Southern Economic Journal*, April 1983, *49*, 975–90.

Freeman, Richard, "Black-White Income Differences: Why Did They Last So Long?," mimeo., University of Chicago, 1972.

Fuller, Wayne and Battese, George, "Estimation of Linear Models With Crossed-Error Structure," *Journal of Econometrics*, May 1974, *2*, 67–78.

Hanushek, Eric, *Education and Race*, Boston: Lexington, 1972.

Harlan, Louis, *Separate and Unequal: Public School Campaigns and Racism in the Southern Seaboard States, 1901–1915.* Chapel Hill: University of North Carolina Press, 1958.

Hausmann, J. A., "Specification Tests in Econometrics," *Econometrica*, November 1978, *46*, 1251–71.

Higgs, Robert, *Competition and Coercion: Blacks in the American Economy, 1865–*

1914, New York: Cambridge University Press, 1977.

Kousser, J. Morgan, "Progressivism—For Middle Class Whites Only: North Carolina Education, 1880–1910," *Journal of Southern History*, May 1980, *46*, 169–94.

Magnum, Charles S., *The Legal Status of the Negro*, Chapel Hill: University of North Carolina Press, 1940.

Margo, Robert A., "'Teachers Salaries in Black and White': The South in 1910," *Explorations in Economic History*, June 1984, *21*, 306–26.

______, (1985a) *Disfranchisement, School Finance, and the Economics of Segregated Schools in the United States South, 1890–1910*, New York: Garland Press, 1985.

______, (1985b) "Educational Achievement in Segregated School Systems: The Effects of 'Separate-but-Equal'," NBER Working Paper No. 1620, May 1985.

______, "Accounting for Racial Differences in School Attendance: The American South, 1900," mimeo., University of Pennsylvania, June 1986.

Smith, James, "Race and Human Capital," *American Economic Review*, September 1984, *74*, 685–98.

Summers, Anita and Wolfe, Barbara, "Do Schools Make a Difference?," *American Economic Review Proceedings*, May 1977, *67*, 639–52.

Welch, Finis, "Education and Racial Discrimination," in Orley Ashenfelter and Albert Rees, eds., *Discrimination in Labor Markets*, Princeton: Princeton University Press, 1973.

Alabama Educational Survey Commission, *Public Education in Alabama*, Washington: American Council on Education, 1945.

State of Alabama, State Superintendent of Public Instruction, *Annual Report*, Montgomery: State Printer, 1920, 1930, 1940.

Tuskegee Institute, *Negro Year Book*, Tuskegee: Negro Year Book Publishing Co. 1930.

U.S. Department of Commerce, *Negro Population, 1790–1915*, Washington: USGPO, 1918.

______, *Fourteenth Census of the United States: 1920, Population*, Vol. III: *Composition and Characteristics of the Population, By States*, Washington: USGPO, 1921.

______, *Fifteenth Census of the United States: 1930, Population*, Vol. III, *Part One*, Washington: USGPO, 1932.

______, *Negroes in the United States, 1920–1932*, Washington: USGPO, 1935.

______, *Sixteenth Census of the United States: 1940, Population*, Vol. II: *Characteristics of the Population, Part One*, and *Housing*, Vol. II: *General Characteristics, Part 2*, Washington: USGPO, 1943.

______, *Census of Population: 1950*, Vol. II: *Characteristics of the Population, Part Two, Alabama*, Washington: USGPO, 1952.

______, *Historical Statistics of the United States: Colonial Times to 1970*, Washington; USGPO, 1975.

Is Price Flexibility Destabilizing?

By ROBERT A. DRISKILL AND STEVEN M. SHEFFRIN*

Sluggish wages and prices are generally the culprits in Keynesian models of unemployment and business fluctuations. Price flexibility in standard fix-price models would restore the economy to full employment. This line of reasoning has often been at the heart of proposals to reform institutions in order to restore flexibility to wages and prices.

There is, however, a strand of macroeconomic thought that questions the wisdom of too much price flexibility. John Maynard Keynes raised the issue in chapter 19 of the *General Theory* by noting that a deflation could raise real interest rates and thereby impede a return to full employment. In his 1975 paper "Keynesian Models of Recession and Depression," James Tobin develops this point in a formal model. Low prices work to move the economy to full employment but falling prices, to the extent that they lead to expectations of deflation, raise the real interest rate (through the Mundell effect) and move the economy away from full employment. Instability is likely to occur if expectations of inflation adjust rapidly to actual inflation and the real interest rate effect is large.

Our historical experience does include significant episodes where either reductions in inflation or actual deflation were accompanied by high real interest rates. Although other factors could be responsible, the experiences of the Great Depression, the Latin American countries in the late 1970's and the United States in the early 1980's all add surface plausibility to the real interest rate deflation link and thus to one aspect of the Keynes-Tobin story.

Recently, Bradford De Long and Lawrence Summers (1984) have argued that the decrease in the variance of output following World War II can be largely attributable to the *decrease* in wage and price flexibility in the postwar era. The reason output fluctuations are smaller today is precisely because the Keynes-Tobin destabilizing mechanism is less operative today.[1]

This paper examines whether increased price flexibility can be destabilizing in a version of John Taylor's contract model (1979, 1980) extended to include real interest rate effects. Taylor's model includes both backward and forward elements in wage-setting behavior and thus permits some rationality in the wage-setting process. We find that even with this limited degree of rationality, increased wage flexibility leads to a decrease in both the variance of output and the variance of prices.

Nicholas Carlozzi and Taylor (1983) introduce the real rate of interest into a staggered contract framework and discuss in general terms and through simulations the effects that changing real rates may have on the system. They do not, however, study the effects of potential instability through increases in wage flexibility or provide any analytical results. They provide an extensive discussion of the implications of alternative policy rules on the stochastic behavior of the economy.

We first add the real interest rate to the standard Taylor model and derive the solution and prove key analytical results. These results are further buttressed by simulations.

*Departments of Economics, Ohio State University, Columbus, OH 43210, and University of California, Davis, CA 95616, respectively. Earlier versions of this paper were presented at Tulane University and the University of New Mexico. Support was provided by the Research Program in Applied Macroeconomics and Macro Policy, University of California-Davis.

[1]Our work on this paper was partly motivated by some of the ideas expressed in De Long and Summers and their suggestion to explore the possibility of Keynes-Tobin instability in a model with rational expectations.

I. Introducing the Real Rate of Interest

To study the possibility of Keynes-Tobin instability, it is necessary to introduce the real interest rate into aggregate demand. First, let output depend on the expected real interest rate:

$$(1) \qquad y_t = -a\{i_t - E_t P_{t+1} + P_t\}.$$

Nominal interest rates are determined in the money market:

$$(2) \qquad m_t - P_t = -\beta i_t \qquad \beta > 0.$$

Finally, let the Fed follow a nominal interest rate rule (which includes as a special case a constant money supply):

$$(3) \qquad m_t^s = \alpha i_t.$$

The interest rate is measured in natural units while all other variables are in logs. Combining (1)–(3) yields the expression for aggregate demand:

$$(4) \qquad y_t = aE_t P_{t+1} - bP_t$$

where $\quad b = a(1 + (1/\beta + \alpha)).$

We assume $\beta + \alpha > 0$, a condition satisfied by a constant money stock, which implies $b > a$.

The aggregate supply side of the model consists of an equation for wage setting and a mark-up equation for pricing. Money wage contracts are set for two periods and depend on a weighted average of wages for workers in the second period of their contract and the wages those workers are expected to negotiate next period as well as a weighted average of current and expected real output. The weights ϕ and $1-\phi$ measure the degree of "backward looking" vis-à-vis "forward looking" in the wage-setting process:

$$(5) \qquad X_t = \phi X_{t-1} + (1-\phi)E_{t-1}X_{t+1} + \gamma[\phi E_{t-1}y_t + (1-\phi)E_{t-1}y_{t+1}] + u_t$$
$$\gamma > 0; \quad 0 \leq \phi \leq 1$$

$E_{t-1}(\cdot)$ is the expectations operator, conditional on information available at time $t-1$, and u_t is a white-noise wage shock. The parameter γ measures the extent of the sensitivity of wages to business cycle conditions. Prices are determined as a mark up over prevailing contract wages in the economy:

$$(6) \qquad P_t = \tfrac{1}{2}(X_t + X_{t-1}).$$

The structural model consisting of equations (4), (5), and (6) involves expectations of variables at time t and $t-1$, current stochastic shocks, current endogenous variables and only one state variable, lagged wages. Leon Wegge (1982) has shown that such a rational expectations structural model has the following solution form for the contract wage:

$$(7) \qquad X_t = \rho X_{t-1} + \pi u_t,$$

where ρ and π are, in general, functions of structural parameters. Combining (7) with the pricing equation (6), we can write prices as the following ARMA(1,1) process:

$$(8) \qquad P_t = \rho P_{t-1} + k(u_t + u_{t-1}) \qquad k = \pi/2.$$

Taking expectations of P_{t+1} at time t, we have

$$(9) \qquad E_t P_{t+1} = \rho P_t + ku_t.$$

Substituting (9) into the expression for aggregate demand and simplifying yields

$$(10) \qquad y_t = -\hat{\theta}P_t + aku_t \qquad \hat{\theta} = b - \rho a.$$

Aside from the error term, this expression for aggregate demand is the same as Taylor's (1979) model which has aggregate demand solely a function of the real money supply and a money supply equation linking the nominal money supply to prices. In this model, however, $\hat{\theta}$ now depends on the (unknown) value of ρ. The final step in the solution is to exploit the similarities between the real rate case and the standard Taylor model to determine the value of ρ.

Treating $\hat{\theta}$ as predetermined for the moment allows us to calculate ρ and π by the standard method of undetermined coefficients:

$$(11)\qquad \rho = \frac{\phi}{\hat{c} - \rho(1-\phi)}; \quad \pi = 1$$

where $\hat{c} = \left(1 + \frac{\hat{\theta}\gamma}{2}\right) \Big/ \left(1 - \frac{\hat{\theta}\gamma}{2}\right).$

However, since $\hat{\theta} = b - \rho a$, there is a second relationship between ρ, and $\hat{c}$, namely:

$$(12)\quad \hat{c} = \left(1 + \frac{\gamma(b-\rho a)}{2}\right) \Big/ \left(1 - \frac{\gamma(b-\rho a)}{2}\right)$$

Equations (11) and (12) determine ρ and $\hat{c}$, for any given values of γ, b, and a. We can now state and prove the following propositions:

PROPOSITION 1: *For any feasible values of γ, b, and a, there exists a unique stable autoregressive parameter ρ for the wage process* (7).[2]

PROOF:

Rearrange (11) as follows:

$$(13)\qquad \hat{c} = \left(\phi + \rho^2(1-\phi)\right)/\rho$$

Equating (12) to (13), cross-multiplying and collecting terms, we have:

$$(14)\quad \frac{2(1-\rho)[\phi - \rho(1-\phi)]}{(1+\rho)[\phi + \rho(1-\phi)]} = \gamma(b - \rho a).$$

Denote the left-hand side of (14) as $f(\rho)$ and the right-hand side as $g(\rho)$. Define $D(\rho) \equiv f(\rho) - g(\rho)$. What we need to show is that $D(\rho)$ equals zero only once over $\rho \in (-1,1)$. We need to prove this proposition for each of two cases.

Case 1: $0 \leq \phi \leq 1/2$, *when* $0 \leq \phi \leq 1/2$, $\phi/(1-\phi) \leq 1$.

[2]Details of the calculations in the proofs are available from the authors.

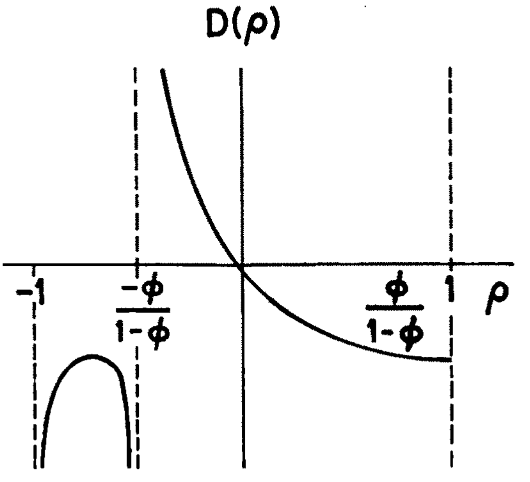


FIGURE 1. THE FUNCTION $D(\rho)$ FOR $0 \leq \phi \leq 1/2$

Straightforward calculation shows that, in this case, $D(\rho)$ is a strictly convex, continuous function for $\rho > -\phi/1-\phi)$ with $D(-\phi/(1-\phi)) = +\infty$ and $D(1) < 0$. For $\rho \in (-1, -\phi/(1-\phi))$, $D(\rho)$ is strictly negative. Hence, over $\rho \in (-1,1)$, $D(\rho)$ equals zero for only one value of ρ. The function $D(\rho)$, for this case, is depicted in Figure 1.

Case 2: $1/2 \leq \phi \leq 1$, *when* $\phi \in [1/2, 1]$, $\phi/(1-\phi) > 1$.

$D(\rho)$ is a continuous, strictly convex function over $\rho \in (-1,1)$ with $D(-1) = +\infty$ and $D(1) < 0$. Hence $D(\rho) = 0$ only once over $\rho \in (-1,1)$. The function $D(\rho)$ for this case is depicted in Figure 2.

We now can prove Proposition 2.

PROPOSITION 2: *Increases in γ, the wage flexibility parameter, decrease the autocorrelation parameter ρ.*

PROOF:

We need to show that $D(\rho)$ shifts down when γ increases. Taking the partial derivative of $D(\rho)$ with respect to γ yields

$$\frac{\partial D(\rho)}{\partial \gamma} = -(b - \rho a) < 0 \quad \text{for } \rho \in (-1,1).$$

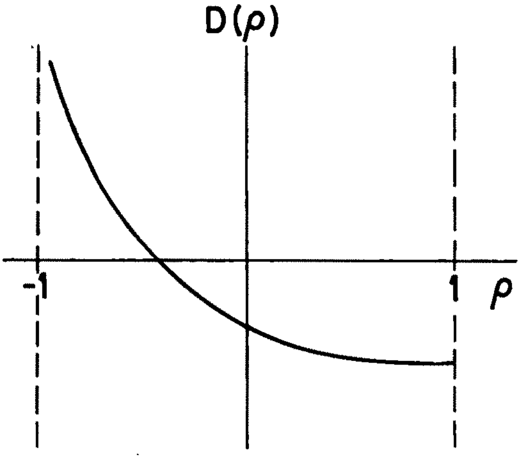


FIGURE 2. THE FUNCTION $D(\rho)$ FOR $1/2 \leq \phi \leq 1$

II. Behavior of the System

We can now examine the lag patterns of prices and output and also the responsiveness of the asymptotic variances of prices and output to changes in the wage flexibility parameter. The incorporation of the real rate makes it possible that an increase in wages and prices arising from a positive realization of the shock u_t would lead to a contemporaneous *increase* in output, which is not possible without the real rate of interest.

By repeated lagging and substituting of (8) and (10), prices and output can be expressed as functions of current and all past shocks:

$$(15a)\quad P_t = \tfrac{1}{2}\{u_t + (\rho+1)u_{t-1} + \rho(\rho+1)u_{t-2} + \rho^2(1+\rho)u_{t-3} + \dots\}$$

$$(15b)\quad y_t = \omega_1 u_t + (\omega_2 + \rho\omega_1) \times \{u_{t-1} + \rho u_{t-2} + \rho^2 u_{t-3} + \dots\}$$

where

$$(16a)\quad \omega_1 = (a(1+\rho) - b)/2;$$

$$(16b)\quad \omega_2 = -(b/2).$$

From (15a), we see that the price level exhibits a "humped" lag pattern. a positive shock u_t pushes the price level up in the current period, then pushes it even higher in the next period, after which time the price level gradually decreases back to its long-run value. This means that a current positive shock increases both the current price level and also creates expected inflation from the current period to the next.

The increase in expected inflation associated with the positive realization of u_t means that the current-period real rate of interest can fall. Hence, output in the current period can increase. In terms of (16a), this means that ω_1 could be positive. This feature of our model is clearly a result of the incorporation of the real rate of interest in aggregate demand.

This "Tobin effect" of an increase in prices creating expected inflation which in turn creates a lower real interest rate can only last one period in our model. After that, the model predicts expected deflation, and the real rate must be above its long-run value. The quantity $\{\omega_2 + \rho\omega_1\}$ is negative, confirming that the effect of a positive wage shock reduces output in subsequent periods, regardless of the first period effect. In Figure 3, we sketch possible lag patterns for output and prices.

The asymptotic variances of prices and outputs can be computed from (15) as

$$(17a)\quad \sigma_p^2 = \sigma_u^2/2(1-\rho)$$

$$(17b)\quad \sigma_y^2 = \{(b-a)^2 + b^2 + \rho a(a-2b)\} \times \sigma_u^2/4(1-\rho).$$

We can now state and prove the following proposition:

PROPOSITION 3: *Increases in γ, the measure of wage flexibility, decrease both price and output asymptotic variability.*

PROOF:

From Proposition 2, we know that increases in γ decrease ρ.

Now,

$$d\sigma_p^2/d\rho = \sigma_u^2/2(1-\rho)^2 > 0.$$

Hence, increases in γ decrease ρ which decreases price variability.

For output variability, we compute:

$$d\sigma_y^2/d\rho = (b-a)^2\sigma_u^2/2(1-\rho)^2 > 0.$$

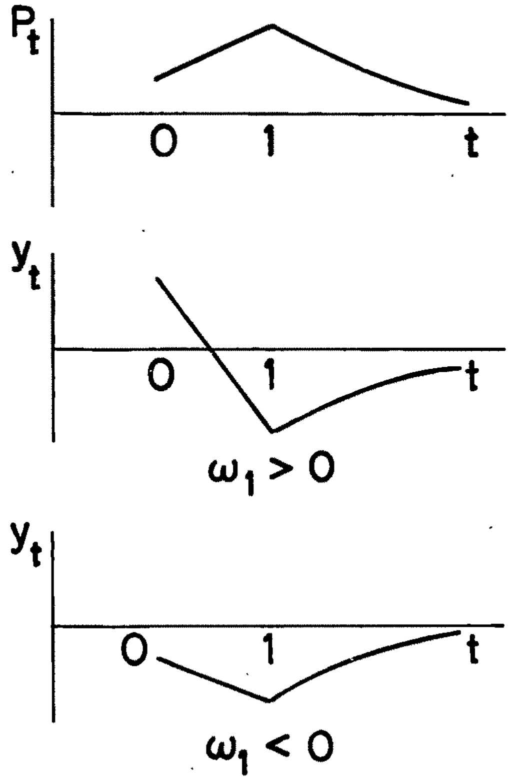


FIGURE 3. LAG PATTERNS FOR P_t AND y_t

Hence, increases in γ decrease ρ which decrease output variability.

III. The Wage-Setting Process with a Larger Information Set

Our model is similar to Taylor's in that expectations in the wage-setting process are conditioned on information available at time $t-1$ rather than at time t. We now investigate the model with the assumption of the larger information set. While analytic comparative-static results are not available, simulations indicate that increases in wage flexibility continue to produce decreased price and output variability.

The only difference between the model we now investigate and the previous one is that expectations in the wage process are conditioned on current information:

$$(18)\quad X_t = \phi X_{t-1} + (1-\phi)E_t X_{t+1} + \gamma[\phi y_t + (1-\phi)E_t y_{t+1}] + u_t.$$

TABLE 1—VARIANCES OF OUTPUT AND PRICES

	a			
γ	.5	.8	1.1	1.4
.1	5.19	3.69	2.70	2.24
	.87	1.35	1.98	2.88
.2	2.99	1.95	1.43	1.10
	.61	.87	1.30	1.82
.3	2.03	1.31	.94	.74
	.46	.66	.97	1.41
.4	1.55	.98	.72	.54
	.38	.53	.81	1.14

Notes: Entries are σ_p^2, σ_y^2, respectively. For the simulations $b/a = 1.5$, $\sigma_u^2 = 1$, and $\phi = .5$.

The solution for the wage equation is now

$$(19)\quad X_t = \rho X_{t-1} + (2\rho(1+\gamma\phi a\pi/2) / [2-\gamma(b-\rho a)]\phi)u_t = \rho X_t + \pi u_t$$

where ρ is identical to ρ in the previous model. The only difference the expanded information set makes is that the coefficient on the current shock term in the autoregressive wage equation is no longer one, but a function of structural parameters. The constraint that π equal the coefficient on u_t in (19) implies the following expression for π:

$$(20)\quad \pi = 2\rho/\phi[2-\gamma b].$$

Asymptotic variances of price and output now become

$$(21a)\quad \sigma_p^2 = \pi^2\sigma_u^2/2(1-\rho);$$

$$(21b)\quad \sigma_y^2 = \hat{\theta}^2\sigma_p^2 + ((1-2\hat{\theta})\pi^2/4)\sigma_u^2.$$

Analytic results on the comparative-statics effects of changes in γ on σ_p^2 and σ_y^2 are not available, but the simulations displayed in Table 1 illustrate that over wide ranges of parameter values, both σ_p^2 and σ_y^2 decrease as γ increases. The expanded information set does not seem to affect the qualitative results of the preceding sections.

IV. Conclusion

Our results limit the possible scope for Keynes-Tobin instability which could arise from the interaction of "too rapid" price and wage adjustment and real interest rate

effects. With economic agents that are partially forward looking, increased wage flexibility only improves the operating characteristics of the system even though the "Tobin" effect is operative in our model. However, in an attempt to obtain analytical solutions, we have only investigated the effects of serially uncorrelated wage disturbances in a two-period contract model and have not explored the effects of serially correlated disturbances, demand shocks, or longer contract length for Keynes-Tobin instability.

REFERENCES

Carlozzi, Nicholas and Taylor, John B., "International Capital Mobility and the Coordination of Monetary Rules," NBER Working Paper No. 1242, 1983.

De Long, J. Bradford and Summers, Lawrence H., "The Changing Cyclical Variability of Economic Activity in the United States," NBER Working Paper No. 1450, 1984.

Keynes, John Maynard, The *General Theory of Employment Interest and Money*, New York: Harcourt, Brace and World, 1965.

Taylor, John B., "Staggered Wage Setting in a Macro Model," *American Economic Review Proceedings*, May 1979, *69*, 108–13.

______, "Aggregate Dynamics and Staggered Contracts," *Journal of Political Economy*, February 1980, *88*, 1–23.

Tobin, James, "Keynesian Models of Recession and Depression," *American Economic Review Proceedings*, May 1975, *65*, 195–202.

Wegge, Leon, "The Canonical Form of the Dynamic Rational Future Expectations Model," mimeo., 1982

On the Limitations of Government Borrowing: A Framework for Empirical Testing

By JAMES D. HAMILTON AND MARJORIE A. FLAVIN*

How long can government budget deficits continue unchecked? This question raises two separate issues. First, are perpetual deficits desirable—are the effects on inflation, investment, and the balance of payments ones that we can live with? Second, are perpetual deficits feasible—even if the government wanted to run a budget deficit forever, is this something it really could do?

If we were talking about the budget plans of a private household, clearly the question of feasibility would be paramount, for we entertain little doubt that households would like to run a permanent deficit if they could get away with it, continually rolling over debt without having to pay anything back and enjoying the associated free lunch. We presume that households do not generally engage in such behavior primarily for the reason of feasibility—no one would be willing to continue lending money to such a household. For this reason, we usually specify that households are subject to the borrowing constraint that the expected present value of expenditures (exclusive of interest payments) not exceed the expected present value of receipts.

The question we pose in this paper is whether governments are subject to an analogous constraint—when a government runs a deficit, is it making an implicit promise to creditors that it will run offsetting surpluses in the future? If governments are subject to this constraint, which we will term the present-value borrowing constraint, the policy of running a permanent deficit (exclusive of interest payments) is infeasible; (though, as we shall see below, a permanent deficit when interest payments are counted as part of the deficit may still be feasible). The question of feasibility of a permanent deficit (exclusive of interest payments) holds profound implications for macroeconomic theory and practice. If governments intend to raise the needed revenues with future tax increases, then government deficits may have no stimulative effect on aggregate demand,[1] but can have significant distortionary effects on private incentives if the future tax increases are large (see Robert Barro, 1984b). On the other hand, if the revenues are to be raised implicitly through money creation, budget deficits can be a principal cause of inflation, as suggested by Thomas Sargent (1982) and Sargent and Neil Wallace (1981).

Whether governments can continually run a budget deficit remains an unsettled theoretical question. If the government borrows at an interest rate that equals or exceeds the economy's growth rate, then a continuing unpaid deficit implies that the debt must

*Department of Economics, University of Virginia, Charlottesville, VA 22901. This paper was written while we were visiting at the University of California-San Diego. We are indebted to Charles Engel, Ron Michener, and John Taylor for helpful comments. Research support from the National Science Foundation under grant SES-8420434 is gratefully acknowledged.

[1]If government spending (G_t), taxes (T_t) and debt (B_t) are related by

$$B_t + \sum_{j=1}^{\infty} (1+r)^{-j} G_{t+j} = \sum_{j=1}^{\infty} (1+r)^{-j} T_{t+j}$$

then any proposal that changes taxes but leaves spending the same must be such that the right-hand side of the above expression remains constant. Thus, permanent income defined by

$$\sum_{j=1}^{\infty} (1+r)^{-j} \left(W_{t+j} L_{t+j} - T_{t+j} \right)$$

would be completely unaffected by any such tax policy. See Robert Barro (1974, 1984a) and Olivier Blanchard (1985) for further discussion.

grow to become an infinite multiple of GNP.[2] Equilibrium models in which investors would continue to buy government debt under such circumstances have proven difficult to develop; see Bennett McCallum (1984) for a clear discussion of the issues. If the real interest rate is less than the growth rate, by contrast, deficits could continue forever without an increase in the ratio of debt to GNP. Theoretical models that seem to allow this possibility have been explored by Willem Buiter (1979), Jonathan Eaton (1981), and Jeffrey Carmichael (1982), though these results have not gone unchallenged (see, for example, John Burbidge, 1983).

In any case, it seems desirable to supplement these theoretical considerations with empirical evidence. David Aschauer (1985) and John Seater and Roberto Mariano (1985), among others, have tested the hypothesis that the government's receipts must equal its expenditures in present-value terms jointly with a permanent income hypothesis, and accepted. Paul Evans (1985) documented the absence of statistical correlation between U.S. budget deficits and interest rates, which he interpreted as evidence in support of this same joint hypothesis. Barro (1984b) tested the hypothesis that the government is subject to the present-value borrowing constraint jointly with the assumption that taxation and deficit policies have historically been optimal, and again accepted. However, to our knowledge there has been no direct empirical test of the present-value borrowing constraint itself.

At first glance, a test of the present-value borrowing constraint might seem straightforward enough. The U.S. government, for example, has run more or less a chronic deficit since 1930, suggesting that permanent deficits are quite feasible and practical. However, this simple argument ignores the potential role of debt retirement through monetization and likewise ignores capital gains on bonds or tangible assets through inflation. Moreover, the official government deficit includes interest payments, whereas we will argue below that the correct magnitude for purposes of testing the present-value borrowing constraint should exclude such payments. Finally, a formal test of whether historical deficits were rationally anticipated and allowance for what might rationally be expected to happen out of sample seems necessary to evaluate this hypothesis adequately.

In this paper we propose an empirical framework for testing the practical limits to public borrowing which addresses these criticisms. We show that the proposition that the government can accumulate ever-growing debt through perpetual deficit financing has a mathematical parallel in the proposition that prices can rise continually in a self-fulfilling speculative bubble. Thus, empirical tests that have been developed for the latter hypothesis may also be fruitfully applied to study the limits of government borrowing.

The conclusions we draw from these tests complement those of Barro (1984a) and Robert Eisner and Paul Pieper (1984), who have noted that while the official budget has registered a chronic deficit, the real value of government debt fell substantially in the postwar period, suggesting that the official accounts have grossly misstated the true fiscal posture of the government. Once an economically reasonable definition of the government budget deficit is adopted, the data seem fully compatible with the assertion that the government budget historically has been balanced in expected present-value terms. Our tests show this conclusion to be reasonably robust with respect to specification of the information on which creditors were basing their forecasts of future surpluses.

[2]Again we are referring to the government deficit exclusive of interest payments. To take a simple example, let output $Q(t)$ grow at the rate $q(Q(t)=Q_0e^{qt})$, and let spending $gQ(t)$ and taxes $xQ(t)$ be fixed multiples of GNP. If r is the cost of borrowing and $B(t)$ is the debt, then debt accumulates according to

$$dB/dt=(g-x)Q(t)+rB(t)$$

implying $$B(t)=\{B_0+tQ_0(g-x)\}e^{qt} \quad \text{if } q=r$$

$$=\{(g-x)/(q-r)\}Q_0(e^{qt}-e^{rt})+B_0e^{rt} \quad \text{otherwise,}$$

which explodes relative to Q_0e^{qt} whenever $g>x$ and $q\leq r$.

Past deficits have been followed by increases in revenues which covered the government's interest obligations, though the implicit "inflation tax" has historically made an important contribution.

Section I provides a formal statement of the present-value borrowing constraint to be tested. Section II briefly discusses some issues of data and measurement. In Section III we present the results of alternative empirical tests of whether the postwar record of budget deficits in the United States could be consistent with the present-value borrowing constraint under rational expectations, with brief conclusions offered in Section IV.

I. The Present-Value Government Borrowing Constraint

Suppose we collected all government debt of a given coupon and maturity into group j. Let $d_{j,t}$ denote the nominal market value of such debt at the end of period t and $\theta_{j,t}$ the total nominal coupon payments between dates $t-1$ and t. We further let P_t denote an aggregate price index of goods in the economy and r the *ex post* real interest rate that is earned on one-period government bonds during an average year.

Suppose no new bonds of type j are issued or redeemed during period t. Then changes in the market value of group j debt can be evaluated using a simple term-structure argument. Define $v_{j,t}$ to be the real excess one-period holding yield of j bonds relative to the average earned on a comparable investment in one-period bonds:

$$(1)\qquad v_{j,t} \equiv \frac{d_{j,t}+\theta_{j,t}}{P_t} - \frac{(1+r)d_{j,t-1}}{P_{t-1}}.$$

If real interest rates were white noise (or constant) and if the expectations theory of the term structure held, then $E_{t-1}v_{j,t}$ would equal zero. In general, positive values of $v_{j,t}$ mean bondholders have made a capital gain on long-term government debt, or that short-term rates are higher than average.

The market value of outstanding government debt will also change due to operations of the Treasury and Federal Reserve. Let T_t denote real tax revenues, G_t real government purchases of goods (excluding interest payments on the debt), and R_t the dollar flow of nominal interest payments made to the public divided by the price level. Thus, the official deficit in constant dollars is measured by

$$(2)\qquad G_t + R_t - T_t,$$

and new debt must be issued in this amount during the year. Changes in the stock of high-powered money (M_t) likewise retire an amount of debt whose market value is[3]

$$(3)\qquad (M_t - M_{t-1})/P_t.$$

Finally, let B_t denote the real market value of debt held by the public ($B_t = \Sigma_j d_{j,t}/P_t$). From (1)–(3), its value is given by

$$(4)\qquad B_t = (1+r)B_{t-1} - \sum_j \theta_{j,t}/P_t + \sum_j v_{j,t} + G_t + R_t - T_t - (M_t - M_{t-1})/P_t + U_{1,t}.$$

In principle, equation (4) might be thought of as an accounting identity. In practice, we must append the error term $U_{1,t}$ since bond purchases or sales towards the beginning of the period would have taken place at market prices closer to those characterizing B_{t-1} rather than B_t, and since measurements of the real market value of net government indebtedness are necessarily imperfect (see Section II below). Again abstracting from this issue of intraperiod timing, note further that $\Sigma_j \theta_{j,t}/P_t + U_{2,t} = R_t$. Thus

$$(5)\qquad B_t = (1+r)B_{t-1} - S_t + V_t,$$

where

$$(6)\qquad S_t \equiv T_t + (M_t - M_{t-1})/P_t - G_t$$

$$V_t \equiv \sum_j v_{j,t} + U_{1,t} + U_{2,t}.$$

[3] We are formally modelling assets held by the government as negative values of $d_{j,t}$, so that foreign exchange transactions by the Federal Reserve are also included in (3). Details of the variables used in the empirical analysis are provided in Section II.

The key points to note about (5) are: (a) it describes the market rather than the par value of government debt, since open market operations by the Federal Reserve retire debt at market value; (b) the measure of the surplus S_t *excludes* interest payments from government spending; and (c) money seigniorage $(M_t - M_{t-1})/P_t$ is added to taxes T_t as a source of revenue for retiring outstanding government debt.

By recursive substitution forwards, equation (5) is seen to imply

$$(7)\quad B_t = \sum_{i=t+1}^{N} \frac{(S_i - V_i)}{(1+r)^{i-t}} + \frac{(1+r)^t B_N}{(1+r)^N}.$$

Equation (5) and its implication (7) cannot be a point of serious controversy, for they do little more than summarize the definitions of monetary and fiscal policy. What is of economic interest (and subject in principle to empirical refutation) is what creditors expect to happen to the second term in (7) as N gets large. Indeed, letting E_t denote the expectations of creditors based on information available at date t, it is clear from (7) that the hypothesis that the government is subject to the present-value borrowing constraint,

$$(8a)\qquad H_0\colon B_t = E_t \sum_{i=t+1}^{\infty} \frac{(S_i - V_i)}{(1+r)^{i-t}},$$

is mathematically equivalent to the restriction that the real supply of bonds held by the public is expected to grow no faster on average than the rate of interest:

$$(8b)\qquad H_0\colon E_t \lim_{N\to\infty} \frac{B_N}{(1+r)^N} = 0.$$

Using the equivalence between (8a) and (8b), we are now in a position to comment in more detail on the precise nature of the present-value borrowing constraint. Note first that condition (8) can be consistent with a permanent government deficit as conventionally measured, that is, inclusive of interest rates. If a constant (interest inclusive) deficit of $-S_t + V_t + rB_{t-1} = k$ were maintained forever, for example, we see from (5) that $B_N = Nk + B_0$ and $\lim_{N\to\infty} B_N/(1+r)^N = 0$. Thus a policy of keeping the interest component of the deficit from rising will ultimately force the government to pay off its debts in present-value terms. On the other hand, a permanent deficit exclusive of interest payments $-S_t + V_t = k$ is not consistent with (8), for then $B_N = k[(1+r)^N - 1]/r + (1+r)^N B_0$ and $\lim_{N\to\infty} B_N/(1+r)^N = k/r + B_0$. Recall McCallum's demonstration that a permanent deficit inclusive of interest payments could be consistent with optimizing behavior by bondholders (and would satisfy our condition (8), whereas a permanent deficit exclusive of interest payments is not consistent with optimizing behavior in his model (and would violate (8)). Finally, we note that condition (8) does not imply that the national debt must eventually be paid off. In fact, (8) is consistent with a constantly increasing stock of debt, as long as the rate of increase is less than the government's borrowing rate. Rather, the question we pose is whether interest on this debt is to be paid with future tax increases or instead with continual issue of new debt.

If (8) represents the null hypothesis that the government budget must be intertemporally balanced, how may we usefully frame the alternative possibility that government deficits (i.e., negative values of S_t) need not be balanced with future surpluses? One interesting class of alternative hypotheses is obtained by assuming that

$$E_t \lim_{N\to\infty} \left[B_N/(1+r)^N\right] = A_0 > 0.$$

Thus, we allow the possibility that a certain annual amount of real government expenditures $r(A_0 - B_0)$ need never be paid for with taxes. From equation (7) we then obtain

$$(9)\quad B_t = E_t \sum_{i=t+1}^{\infty} \frac{(S_i - V_i)}{(1+r)^{i-t}} + A_0(1+r)^t$$

as a general class of solutions to (5). The hypothesis H_0 that the government budget must be balanced in present-value

TABLE 1—ILLUSTRATIVE SUMMARY OF ADJUSTMENTS TO OFFICIALLY REPORTED SURPLUS FOR FISCAL YEAR 1974 AND FOR DEBT AT END OF FISCAL YEAR 1974
(Millions of current dollars)

Surplus:	
Officially Reported Surplus	−3,460
Plus: interest	+28,072
Minus: interest paid to government trust funds	−6,583
Minus: deposit of Fed earnings	−4,845
Plus: change in agency securities	+903
Equals: Measure of surplus in which interest payments are excluded	+14,087
Plus: net capital gains on gold stock	+3,250
Plus: money seigniorage	+11,331
Equals: true surplus (current dollars)	+28,668
Divided by: consumer price index	÷1.469
Equals: true surplus (1967 dollars)	+19,515
Debt:	
Officially Reported Debt	474,235
Minus: investments in government accounts	−140,194
Equals: par value of Treasury debt	334,041
Multiplied by: market-par ratio	0.951
Equals: market value of Treasury debt	317,740
Minus: currency in circulation	−73,833
Minus: member bank reserves	−30,086
Equals: market value of net interest-bearing debt held by public	213,821
Minus: Treasury operating balance	−9,159
Minus: market value of gold holdings	−39,951
Equals: adjusted government debt (current dollars)	164,711
Divided by: consumer price index	÷1.469
Equals: adjusted government debt (1967 dollars)	112,124

terms holds true if and only if $A_0 = 0$ in equation (9).

Equation (9) is mathematically equivalent to the models of self-fulfilling fads or speculative bubbles first explored by Robert Flood and Peter Garber (1980). We accordingly propose that such tests might also be fruitfully applied to understanding the limits of government borrowing. The next section discusses the data on which such tests might be based, while Section III summarizes our results.

II. Issues of Data and Measurement

This section, inspired in part by Barro (1984a) and Eisner and Pieper, briefly discusses how the theoretical magnitudes appearing in our equation (9) are related to the budget figures actually reported in the United States. Complete details are provided in a data appendix available from the authors on request.

A. *Interest Payments*

The theoretical measure of government spending (G_t) in the above derivation excludes outlays for interest payments. Correcting the officially reported surplus ($\hat{T}_t - \hat{G}_t$) to exclude interest payments from G_t requires three steps. 1) Add to ($\hat{T}_t - \hat{G}_t$) total interest paid by the government. 2) For data prior to the accounting change in 1982, avoid double counting by subtracting back out that part of this Treasury interest that was paid directly into government trust funds such as Social Security. 3) Likewise subtract the deposit of Federal Reserve earnings which are already included in $\hat{T}_t$ under miscellaneous receipts. Sample calculations for corrections to the deficit and debt are provided in Table 1; the actual series used in our empirical work are reported in Table 2.

Some debt is issued by various federal agencies in addition to that issued by the Treasury. Outlays for the service of such

TABLE 2—ADJUSTED VALUES FOR SURPLUS AND DEBT
(Millions of 1967 dollars)

Fiscal Year	Adjusted Surplus for Fiscal Year	Adjusted Debt for End of Fiscal Year
1960	+8,533	167,954
1961	+921	170,826
1962	+2,117	176,187
1963	+3,697	177,944
1964	+3,968	176,036
1965	+9,604	173,319
1966	+11,514	160,103
1967	+7,997	157,441
1968	−3,262	162,227
1969	+14,142	144,302
1970	+6,529	136,877
1971	−3,631	145,977
1972	+1,812	151,836
1973	+13,451	136,272
1974	+19,515	112,124
1975	−13,152	134,673
1976[a]	−35,480	185,230
1977	−4,162	191,707
1978	+3,022	183,956
1979	+28,211	148,641
1980	+20,506	121,432
1981	−27,747	145,895
1982	−22,147	209,445
1983	−38,336	265,164
1984	−28,675	303,205

[a]Data for fiscal year 1976 include the transition quarter (July 1, 1976 through September 30, 1976).

debt are presently included as expenditures of that agency, and should be subtracted from $\hat{G}_t$ to arrive at our measure of G_t. Unfortunately, such data are not readily available. Since agency debt is small (for example, \$12 B in 1974 compared with \$475 B in Treasury debt), little harm can come from *assuming* that the present-value relation holds for agency debt, in which case we could approximate the present value of later agency interest payments by the current market value of new agency debt issue. Thus, we subtract new issue of agency securities from $\hat{G}_t$, and do not count the market value of agency securities in our measure of public holdings of government debt, B_t.[4]

B. *Trust Funds*

Some might argue that trust fund holdings are to be used against the government's liabilities implied by future Social Security benefit payments. It seems to us that this is an inaccurate interpretation. Such programs are not a current liability in the sense that they can be associated with any concrete number. Rather, they represent the outcome of an uncertain political process, and the correct way to represent this "liability" is by the discounted cash flow of an entry on current account rather than any dubious imputation to capital account. For this reason, we follow the official accounts in registering net Social Security inflows or outflows on the deficit account, but differ from the official accounts on the debt account. The correct measure subtracts that money which is owed from one branch of government to another.

C. *Off-Budget Items*

Starting in 1971, the activities of certain agencies such as the Postal Service Fund and Rural Electrification and Telephone Revolving Fund are characterized as "off-budget." Any deficits of such operations require the issue of Treasury bonds, though they do not count in the officially measured deficits of the U.S. government.

Suppose that these funds were indeed used primarily to issue market-interest loans to the private sector. Imagine the agency issuing a \$1 loan financed through a \$1 sale of Treasury bonds, and so running an off-budget \$1 deficit for that year. In the following year, the agency receives \$$r$ as interest payments from the private sector, but the Treasury pays \$$r$ back to the public as interest on the T bond. For this year, the official budget and off-budget items would accordingly sum to a zero net deficit. Thus, the

[4]To take a simple example, suppose that in year 1 spending exceeds taxes by \$1, with the shortfall made up by \$1 issue of new agency debt. In all subsequent years, interest spending is \$$r$ and taxes are correspondingly higher by \$$r$. In the accounts as actually reported, this policy would be associated with a \$1 deficit in year 1 and no surplus in subsequent years. In our proposed measure, by contrast, the deficit is zero in all years; i.e., agency debt has no effect on the present-value calculation for Treasury debt, as it should not in this case.

present-value budget across the two years would register a deficit if one combined the official budget and off-budget items into a single account, whereas the operation itself is clearly fiscally neutral—the government issued a $1 Treasury bond but acquired a $1 private bond, and has simply swapped like assets with the public. A correct measure would be obtained in this case if we adopt our convention of excluding Treasury interest payments from G_t and simply ignore the off-budget surplus or deficit altogether.

Of course, these programs are not pure market loans but in fact have a substantial subsidy aspect. Our only justification for ignoring this is that it is difficult to quantify and presumably small relative to the complete budget.

D. *Net Government Indebtedness*

Eisner and Pieper have begun the difficult task of quantifying the market value of various tangible assets owned by the government. The question for purposes of the present study is, do government bondholders believe that future interest payments will really be met through sale of such assets, rather than by more conventional means such as tax revenues or monetization? Our own view is that, for the vast majority of these assets, the promise of substantial liquidation of government tangible assets is not a politically credible backing for U.S. Treasury debt.

One important exception is the government's gold holdings. There is abundant historical evidence that governments willingly draw down or deplete these stocks in the wake of fiscal crises; this indeed is presumably the primary purpose of holding such stocks in the first place. Let Au_t denote the government's gold holdings in ounces, P_t^{Au} the price per ounce of gold, and D_t the nominal value of debt, all measured at the beginning of period t. Let P_tS_t denote the correctly measured surplus during period t (excluding interest payments from spending but making no correction for gold flows), and let i_t be the one-period interest rate, which for simplicity we assume is the same for all bonds. If at the end of period t the government sells off some amount of gold $(Au_t - Au_{t+1} > 0)$ at price P_{t+1}^{Au} and uses the proceeds to retire debt, then next period's debt will be given by

$$D_{t+1} = (1+i_t)D_t - P_tS_t - (Au_t - Au_{t+1})P_{t+1}^{Au}$$

which can be written as

$$(D_{t+1} - P_{t+1}^{Au}Au_{t+1}) = (1+i_t)(D_t - P_t^{Au}Au_t) + (i_t - \pi_t^{Au})P_t^{Au}Au_t - P_tS_t$$

where $\pi_t^{Au} \equiv (P_{t+1}^{Au} - P_t^{Au})/P_t^{Au}$. Thus, if we define true government indebtedness as the stock of Treasury debt held by the public less the current market value of the government's gold holdings (i.e., as $D_t - P_t^{Au}Au_t$), the equation governing the evolution of true government indebtedness is obtained by subtracting $(i_t - \pi_t^{Au})P_t^{Au}Au_t$ from the officially measured surplus $(\hat{T}_t - \hat{G}_t)$.

For fiscal year 1974, gold prices increased by 17.16 percent, whereas the nominal interest rate on one-year government bonds was 7.56 percent. Based on a market value of the government's gold holdings of \$33.8 B, a sum of $(.1716 - .0756)(33.8) = \$3.2$ B should be added to the surplus for fiscal year 1974 to represent capital gains from gold.

We also need to subtract liquid assets held by the Federal Reserve from our measure of the government debt (see fn. 3). Since these are all carried on the books at par value, the simplest way to construct the correct measure of the market value of these assets is from the liabilities side, namely, by subtracting high-powered money from outstanding government debt to get a measure of net interest-bearing debt held by the public. The Treasury operating cash balance must also be subtracted to arrive at the figure B_t appearing in equation (5).

Painstaking calculations of the true market value of government debt based on actual market quotations of outstanding securities have been updated by W. Michael Cox (1985). At the end of fiscal year 1974, outstanding government debt was trading at a market value of only 95 percent of its par value.

E. *Money Seigniorage*

When the Federal Reserve acquires a Treasury bond, future interest payments on that bond accrue to the Fed and are counted as part of tax revenues under "Miscellaneous Receipts" in the official budget. Since the present value of this tax benefit is equal to the Fed's initial cost of the bond, seigniorage is in this sense theoretically already included in the budget. In practice, however, any finite-sample estimate of the discounted value of these miscellaneous receipts must be less than the discounted value of $M_{t+i} - M_{t+i-1}$, because open market purchases towards the end of the sample period have not yet been amortized. For this reason, we exclude deposits of Fed interest from our measure of T_t and include the seigniorage measure $(M_t - M_{t-1})$, where M_t denotes high-powered money (the sum of currency in circulation plus reserves of member banks).

III. Empirical Tests

We are interested in the question of whether the U.S. government's creditors could rationally expect that the government budget would be balanced in present-value terms. In the light of the discussion of Section I, we state this hypothesis as the restriction $A_0 = 0$ in the formulation

$$(10) \quad B_t = A_0(1+r)^t + E_t \sum_{j=1}^{\infty} (1+r)^{-j} S_{t+j} + n_t$$

where B_t and S_t are the adjusted debt and surplus series reported in Table 2 and n_t is a regression disturbance term reflecting expected changes in real short-term interest rates, the term structure of long rates, and measurement error. The operator E_t denotes the expectations of creditors, which we assume are formed rationally.

Equation (10) is mathematically equivalent to the model proposed by Flood and Garber for studying self-fulfilling hyperinflations. However, Hamilton and Charles Whiteman (1985) expanded on the caveat stated by Flood and Garber that their technique implicitly imposes strong restrictions on the variables used by agents in forming expectations E_t and on the dynamics allowed for n_t. Behzad Diba and Herschel Grossman (1984) and Hamilton and Whiteman suggested that a more general test should first be considered which is more robust with respect to such restrictions. In particular, for *any* stationary process for $(n_t, E_t\Sigma_{j=1}^{\infty}(1+r)^{-j}S_{t+j})$, when $A_0 = 0$, B_t will be stationary, whereas for $A_0 > 0$, B_t will not be stationary. We accordingly initially examine two simple tests.

A. *Dickey-Fuller Test for Unit Roots*

David Dickey and Wayne Fuller (1979, p. 431) suggested the following test of the null hypothesis that a series z_t is nonstationary with unit roots. Estimate

$$z_t - z_{t-1} = \psi_0 + \psi_1 z_{t-1} + \psi_2(z_{t-1} - z_{t-2}) + e_{z,t}$$

by ordinary least squares, and calculate $\hat{\psi}_1/\hat{\sigma}_1$ where $\hat{\sigma}_1$ is the *OLS* standard error for $\hat{\psi}_1$. The null hypothesis (nonstationarity) says this statistic should be zero; the alternative (stationarity) says less than zero.

Using the data in Table 2 we estimated the following equation by *OLS* for $t = 1962$ to 1984 (standard errors in parentheses):

$$S_t - S_{t-1} = \underset{(3.27)}{-0.53} - \underset{(0.24)}{0.70}S_{t-1} + \underset{(0.24)}{0.38}(S_{t-1} - S_{t-2}) + \hat{e}_{S,t}$$

$$B_t - B_{t-1} = \underset{(28.13)}{79.63} - \underset{(0.17)}{0.48}B_{t-1} + \underset{(0.22)}{1.02}(B_{t-1} - B_{t-2}) + \hat{e}_{B,t}$$

The Dickey-Fuller test statistics are -2.92 in the case of the surplus and -2.82 in the case of debt, to be compared with a 5 percent critical value of -3.00 and 10 percent value of -2.63 reported in Fuller (1976, Table 8.5.2, p. 373). The data thus favor

rejection of the null hypothesis of nonstationarity in both cases; that is, the data seem fully compatible with the assertion that investors rationally expected the budget to be balanced in present-value terms.

B. *Generalized Flood-Garber Test*

If expectations of future surpluses are conditioned in part on past surpluses and if we include lagged debt to eliminate the serial correlation of the resulting error term, then equation (10) takes the form

$$B_t = c_0 + A_0(1+r)^t + c_1 B_{t-1} + \dots + c_p B_{t-p} + b_0 S_t + b_1 S_{t-1} + \dots + b_{p-1} S_{t-p+1} + \varepsilon_t. \tag{11}$$

(Here ε_t is the residual from a projection of $[E_t\Sigma_{j=1}^{\infty}(1+r)^{-j}S_{t+j}+n_t]$ on $[S_t, S_{t-1}, \dots, S_{t-p+1}, B_{t-1}, B_{t-2}, \dots, B_{t-p}$, and a constant].) If one were willing to impose stronger restrictions on the dynamics of n_t and on the information set used by creditors, then one would want to estimate (11) jointly with cross-equation restrictions on the process followed by S_t, as in Flood and Garber's study of money demand. In the absence of such restrictions, Hamilton and Whiteman showed that the coefficients c_j, b_j are unrestricted, and a more general test of H_0 is obtained by simple *OLS* estimation[5] of (11):

$$B_t = \underset{(26.40)}{48.41} - \underset{(21.29)}{22.68}\,(1+r)^t + \underset{(0.21)}{0.69}\,B_{t-1} + \underset{(0.24)}{0.20}\,B_{t-2} - \underset{(0.13)}{1.30}\,S_t - \underset{(0.31)}{0.63}\,S_{t-1} + \hat{\varepsilon}_t.$$

We took $r = 0.0112$, the average *ex post* real rate over 1960–84. This equation clearly yields no indication that government debt tends to be growing at rate r; the coefficient $\hat{A}_0$ is statistically insignificant, and, if anything, negative in sign.

While the two tests employed above are more general than most of those appearing in the literature on speculative price bubbles, the motivating assumption that $\{E_t\Sigma_{j=1}^{\infty}(1+r)^{-j}S_{t+j}\}$ follows a stationary process is still not completely general. In particular, if one admits the possibility of out-of-sample changes in regime, the parameter A_0 in equation (10) would be unidentified (see Hamilton, forthcoming, and the references therein). One might further want to admit the possibility of a change in regime in which the government budget had been expected to be balanced in present-value terms up until some date t and only after that date was a permanent deficit introduced, in which case the corresponding "bubble" term would be zero up until date t and $A_t(1+r)^{\tau-t}$ for $\tau > t$.

One obvious candidate for a possible change in regime in both the $\{S_t\}$ series and the bubble term $A_0(1+r)^t$ is the inauguration of Ronald Reagan as president in 1981. As a general test of this possibility, one can imagine allowing all of the parameters in (11) to take on different values before and after 1981, and seeing whether the post-1981 $\hat{A}_0$ is still zero. Unfortunately, there are insufficient degrees of freedom to carry out such a test; if one believes that a change in regime occurred in 1981, it is impossible to determine statistically whether the change represents a new time series process for $\{S_t\}$ that is still consistent with (8), or instead represents a change to a nonzero value for the bubble term A_0. Distinguishing between these possibilities must await additional data. What one can say, however, is that the casual impression afforded by the official deficit series—that the government has been running a permanent deficit for 25 years in complete disregard of the present-value borrowing constraint—is not supported by a closer inspection of the data. For the sample taken as a whole, the data appear quite consistent with the assertion that the govern-

[5]Flood and Garber note in their fn. 18 (p. 754) that caution must be exercised in interpreting the usual t-test in this application. An added complication for our application is that under some specifications of the alternative hypothesis A_0 could be regarded as a random variable. A related issue arises in interpreting Flood and Garber's results if it were thought that some random economic event set off the speculative bubble. Our Dickey-Fuller test is likewise not without problems; if (8) fails, the nonstationary root is not unity as the Dickey-Fuller tests assume but rather $(1+r)$.

ment has historically operated subject to the constraint that expenditures not exceed receipts in expected present-value terms.

C. *Restricted Flood-Garber Test*

The test actually used by Flood and Garber (1980) would be valid for our application only under the further restrictions that n_t follows a white-noise process

$$(12) \qquad n_t = k + \varepsilon_{1t},$$

and that creditors' expectations of future surpluses are based solely on realizations of past surpluses.[6] A straightforward manipulation of the formula derived by Lars Hansen and Sargent (1981, p. 99) yields that for

$$(13) \quad S_t = k_2 + a_1 S_{t-1} + a_2 S_{t-2} + a_3 S_{t-3} + \varepsilon_{2t},$$

then

$$(14) \quad \hat{E}_t \sum_{j=1}^{\infty} b^j S_{t+j} = \frac{bk_2}{(1-b)(1-a_1 b - a_2 b^2 - a_3 b^3)} + \frac{(a_1 b + a_2 b^2 + a_3 b^3) S_t}{(1-a_1 b - a_2 b^2 - a_3 b^3)} + \frac{(a_2 b + a_3 b^2) S_{t-1}}{(1-a_1 b - a_2 b^2 - a_3 b^3)} + \frac{(a_3 b) S_{t-2}}{(1-a_1 b - a_2 b^2 - a_3 b^3)},$$

[6]The second assumption in particular is admittedly unrealistic. Indeed, it can be shown that expectations *must* be based on additional information besides S_t (namely, on the exogenous shocks to which the endogenous policy variable S_t responds) if equation (8) is to hold, because the forecast errors $(E_t - E_{t-1})S_{t+j}$ cannot be fundamental for S_t. It nevertheless seems of interest to see how good an approximation one gets to the data by ignoring this difference between the true forecasts of creditors $(E_t S_{t+j})$ and our econometric forecasts $(\hat{E}_t S_{t+j})$ based on a univariate autoregression for S_t.

TABLE 3—ESTIMATES OF PARAMETERS IN EQUATIONS (13) AND (15)

Parameter	Estimate	Standard Error
A_0	-61.52	(58.20)
k_1	241.51	(68.87)
k_2	0.90	(3.83)
a_1	0.15	(0.19)
a_2	-0.47	(0.22)
a_3	-0.51	(0.20)

where $\hat{E}_t(Y) \equiv E(Y|S_t, S_{t-1}, S_{t-2}, \ldots)$ and $b \equiv 1/(1+r)$. Substituting (12) and (14) into (10) yields

$$(15) \quad B_t = A_0(1+r)^t + k_1 + \frac{(a_1 b + a_2 b^2 + a_3 b^3) S_t}{(1-a_1 b - a_2 b^2 - a_3 b^3)} + \frac{(a_2 b + a_3 b^2) S_{t-1}}{(1-a_1 b - a_2 b^2 - a_3 b^3)} + \frac{(a_3 b) S_{t-2}}{(1-a_1 b - a_2 b^2 - a_3 b^3)} + \varepsilon_{1t}.$$

We estimated equations (13) and (15) jointly by nonlinear least squares for $t = 1963$ to 1984. Parameter estimates and standard errors are summarized in Table 3.

As in the less restrictive tests above, we note that there seems to be no role whatever for the bubble term; $\hat{A}_0$ is statistically insignificant, and, if anything, negative. The assumption that bondholders rationally expected the debt to be paid back in present-value terms fits the data better than the assumption that debt has simply accumulated with an ever-growing interest load. Moreover, for these parameter estimates, equation (15) has an R^2 of 0.53; that is, more than half of the observed variance in the market value of real government debt

could be explained from a rational expectations forecast of the discounted value of future surpluses. Note such high explanatory power is achieved despite the fact that all of the other parameters that characterize the dynamics of outstanding debt in equation (15) are tied down by the univariate process for surpluses (13), and such parameters appear in (15) only to the extent that they could characterize rational expectations forecasts of future surpluses. The high R^2 is also achieved despite the omission of all of the additional variables besides past surpluses that would be used by agents to forecast future surpluses.

Of course, to achieve such a fit to the data, the regression is forced to fit strongly negative coefficients in the autoregressive process for surpluses at two- and three-year lags, as the values in Table 3 indicate. Bondholders must assume that big deficits typically only last a year or two, and will later be balanced out with surpluses. There is indeed moderate support for this position in a completely unconstrained *OLS* estimate of the latter regression:

$$\begin{aligned} S_t = \underset{(3.51)}{-0.30} + \underset{(0.23)}{0.62S_{t-1}} \\ \underset{(0.28)}{-0.30S_{t-2}} - \underset{(0.27)}{0.20S_{t-3}} + \hat{\varepsilon}_{2t} \end{aligned}$$

The restricted rational expectations estimates of Table 3 simply exaggerate this feature in the data. Overall, then, the present-value hypothesis seems to hold up quite well.

IV. Conclusions

In this paper we have examined the proposition that in order to be able to issue interest-bearing debt, a government must promise to balance its budget in expected present-value terms. We suggested a battery of empirical tests of this proposition, some of which are quite robust with respect to assumptions about the dynamics of variables that are seen by agents but not the econometrician, and others which are highly restrictive. The conclusion from all our tests, however, is the same—the proposition that the government must promise creditors that it will balance the budget in expected present-value terms seems largely consistent with postwar U.S. data.

This result might seem surprising since the official budget accounts register essentially uninterrupted deficits for the United States from 1960 to 1981. However, the real value of government debt held by the public actually fell during this period, indicating that the continuing reported deficits grossly misstated the true fiscal posture of the government. We suggested an alternative measure of the government deficit that takes into account revenues from monetization and capital gains on gold but excludes interest payments. From the time-series properties of the adjusted deficit series, one can construct a rational expectations forecast of the present value of future government budget surpluses. Such a forecast series can account for 53 percent of the observed variance of real government debt under the assumption that the government budget must be balanced in present-value terms.

If our conclusion on the limitations of government borrowing is correct, then the prevailing sentiment in Washington that current deficits can continue forever is wrong; the adjusted deficit series must soon turn to surplus. One policy change that could turn the adjusted series to surplus would be a resurgence of money growth.

REFERENCES

Aschauer, David A., "Fiscal Policy and Aggregate Demand," *American Economic Review*, March 1985, *75*, 117–27.

Barro, Robert J., "Are Government Bonds Net Wealth?," *Journal of Political Economy*, November/December 1974, *82*, 1095–117.

______, (1984a) *Macroeconomics*, New York: Wiley & Sons, 1984.

______, (1984b) "The Behavior of U.S. Deficits," NBER Working Paper No. 1309, March 1984.

Blanchard, Olivier J., "Debt, Deficits, and Finite Horizons," *Journal of Political Economy*, April 1985, *93*, 223–47.

Buiter, Willem H., "Government Finance in an Overlapping Generations Model with Gifts and Bequests," in George M. von Furstenberg, ed., *Social Security Versus Private Saving*, Cambridge: Ballinger, 1979.

Burbidge, John B., "Government Debt in an Overlapping-Generations Model with Bequests and Gifts," *American Economic Review*, March 1983, *73*, 222–27.

Carmichael, Jeffrey, "On Barro's Theorem of Debt Neutrality: The Irrelevance of Net Wealth," *American Economic Review*, March 1982, *72*, 202–13.

Cox, W. Michael, "The Behavior of Treasury Securities; Monthly, 1942–1984," Research Paper No. 8501, Federal Reserve Bank of Dallas, February 1985.

Diba, Behzad T. and Grossman, Herschel I., "Rational Bubbles in the Price of Gold," NBER Working Paper No. 1300, 1984.

Dickey, David A. and Fuller, Wayne A., "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association*, June 1979, *74*, 427–31.

Eaton, Jonathan, "Fiscal Policy, Inflation and the Accumulation of Risky Capital," *Review of Economic Studies*, July 1981, *48*, 434–45.

Eisner, Robert and Pieper, Paul J., "A New View of the Federal Debt and Budget Deficits," *American Economic Review*, March 1984, *74*, 11–29.

Evans, Paul, "Do Large Deficits Produce High Interest Rates?," *American Economic Review*, March 1985, *75*, 68–87.

Flood, Robert P. and Garber, Peter M., "Market Fundamentals Versus Price Level Bubbles: The First Tests," *Journal of Political Economy*, August 1980, *88*, 745–70.

Fuller, Wayne A., *Introduction to Statistical Time Series*, New York: Wiley & Sons, 1976.

Hamilton, James D., "On Testing for Self-Fulfilling Speculative Price Bubbles," *International Economic Review*, forthcoming.

______ **and Whiteman, Charles H.,** "The Observable Implications of Self-Fulfilling Expectations," *Journal of Monetary Economics*, November 1985, *16*, 353–73.

Hansen, Lars P. and Sargent, Thomas J., "Formulating and Estimating Dynamic Linear Rational Expectations Models," in Robert E. Lucas Jr. and Thomas J. Sargent, eds., *Rational Expectations and Econometric Practice*, Minneapolis: University of Minnesota Press, 1981.

McCallum, Bennett T., "Are Bond-Financed Deficits Inflationary? A Ricardian Analysis," *Journal of Political Economy*, February 1984, *92*, 123–35.

Sargent, Thomas J., "The Ends of Four Big Inflations," in Robert E. Hall, ed., *Inflation: Causes and Effects*, Chicago: University of Chicago Press, 1982.

______ **and Wallace, Neil,** "Some Unpleasant Monetarist Arithmetic," Federal Reserve Bank of Minneapolis *Quarterly Review*, Fall 1981, *5*, 1–17.

Seater, John J. and Mariano, Roberto S., "New Tests of the Life Cycle and Tax Discounting Hypotheses," *Journal of Monetary Economics*, March 1985, *15*, 195–215.

Tax Holidays as Signals

By ERIC W. BOND AND LARRY SAMUELSON*

When establishing new plants overseas, multinational firms are often offered substantial investment incentives by host countries. Examples of the types of subsidies offered by host countries include reduced tax rates in the early years of operation, cash grants, subsidized loans, and labor training grants. These subsidies have become a widespread practice in both developed and developing countries, and play an important role in efforts to secure foreign investment. In the Common Market, for example, regional incentive programs have become so prevalent that maximum incentive levels have been imposed that limit the size of subsidies offered.[1] Investment incentive programs have also become important parts of successful development programs in Taiwan, Korea, and Puerto Rico. Within the United States, investment incentives to attract firms have become an integral part of the economic programs of many states.

One common feature of these programs is that the benefits to firms are concentrated in the early years of operation. The archetype is the tax holiday, in which the firm receives a reduced tax rate for a fixed period of time but pays taxes at a higher rate when the holidays expires. This practice is puzzling. If the firm and the country have similar discount rates, why does the reduction in tax rates invariably take this form rather than a uniform tax reduction with equivalent present value?

A bargaining-based explanation of tax holidays has been offered by Chris Doyle and Sweder van Wijnbergen (1984). They observe that once a firm has entered a country and incurred fixed costs, the bargaining power of the country increases. The relative attractiveness to the firm of other locations has decreased, since host-country sunk costs have been paid, and the country can exploit this lock-in effect to increase tax rates. In a multiperiod model, the result of the bargaining process will then be a tax rate that increases over time, reflecting the increased bargaining power of the country. This model offers many insights. However, it does not examine a potentially important aspect of tax holidays, their role as a signal.

Our analysis begins with the presumption that a firm is uncertain as to the productivity of the country in which it will potentially locate. A tax holiday (as opposed to uniform tax rates) may play the role of a signal in this model because it potentially allows high-productivity countries to distinguish themselves. A tax holiday does this by calling for subsidies in early periods, in which a country's productivity is unknown. Tax payments occur in subsequent periods, in which the country exploits the fixed investment made by the firm. The firm will tolerate relatively high subsequent tax rates in a high-productivity country, allowing the latter to recoup its initial subsidies. However, the firm will abandon a low-productivity country rather than pay such rates, preventing the low-productivity country from recovering its initial subsidies and profitably offering a similar tax holiday. The high-productivity country can then use tax holidays to identify itself, inducing firms to enter at higher tax rates than would be the case without such

*Department of Economics, Pennsylvania State University, University Park, PA 16802. The helpful comments of Barry W. Ickes are gratefully acknowledged, as is financial support from the National Science Foundation. We are especially appreciative to Chris Doyle for detailed comments that made a significant contribution to the final version of the paper, particularly to the formulation of the extensive form of the model.

[1] Douglas Yuill and Kevin Allen (1981) survey nine regional incentive programs in Europe and find that all countries offer capital grants and interest rate subsidies, four offer tax concessions, and four offer labor-related subsidies. The legal maximum on grants imposed by the Common Market is 60 percent of the cost of the project. Bond and Stephen Guisinger (1985) found that in some industries in Ireland, as much as 40 percent of the protection received by firms is provided by subsidies of this type.

identification. Tax holidays are thus an optimal means by which to convey information.

Two results distinguish our model of tax holidays from the conventional analysis. First, the conventional model requires the presence of fixed cost in order to generate a tax holiday. We find that if the source country is sufficiently attractive relative to the host country, a tax holiday will appear in spite of the absence of any fixed cost. Secondly, we find that the presence of uncertainty can lead to a stricter form of tax holiday (in which the first-period tax rate is negative) than would occur in the conventional model. It is accordingly clear that the potential for tax holidays to act as signals has important implications.

I. The Model

We consider a single firm and country, referred to as a foreign country, who bargain over the tax rates to be paid by the firm if it establishes a plant in the country. The firm is a monopolist serving a world market, and must decide whether to locate the plant in its home country or in a foreign country. An initial investment of fixed cost K is required if the plant is located in the foreign country. This allows the firm to earn a flow of quasi rents per period, R, equal to the difference between the revenue received from sales and the cost of locally purchased variable factors. The quasi rents earned by the firm if it enters the foreign country will depend upon the quality of the country's labor force, infrastructure, or economic and political stability. We assume that the country is one of two possible types, termed high-productivity (H) and low-productivity (L) countries. Quasi rents per period are R_H in the H country and R_L ($< R_H$) in the L country. The local characteristics that determine R can presumably be ascertained by the country, and we accordingly assume that countries know their own productivities. In contrast, the firm is unlikely to be able to collect sufficient information about the country to learn its productivity with certainty. We let p be the (possibly subjective) probability assigned by the firm to the event that the country is an H country, with probability $1-p$ that the country is an L country. We denote expected quasi rents when country type is not known by $R_P = pR_H + (1-p)R_L$. Upon entering the country and undertaking production for one period, the firm learns the productivity of the country with certainty. The life of the investment is assumed to be two periods.

The firm and country share a common discount factor of D. We presume that the country is a tax revenue maximizer. If the firm does not enter a country, the country earns tax revenues of zero, and the firm establishes the plant in its home country. We assume that the profits (net of fixed costs) of the latter alternative are not subject to uncertainty and are common knowledge. The value of this outside alternative is given by $\bar{V}$. We then have $\bar{V} = (1-t_d)R_d + D(1-t_d)R_d - K_d$, where t_d, R_d, and K_d are the domestic tax rates (assumed to be constant across periods), quasi rents, and entry costs. After the first period has passed, the firm retains the option of abandoning its venture in the foreign country and returning to the home country to produce. This yields a known period-2 profit (net of fixed costs) with period-2 present value of $\tilde{V}$, where $\tilde{V} = (l-t_d)R_d - K_d$. The difference between $\bar{V}$ and $\tilde{V}$ is then that the former gives the period-1 present value of entering the domestic market for two periods; the latter the period-2 present value of entering for period 2 only. We assume that $\tilde{V} > 0$, so that the firm will leave the foreign country after one period if the returns are sufficiently low. Figure 1 illustrates the extensive form of this game.

Since the firm has an outside option of $\tilde{V}$ in period 2, the firm will remain in the country only if $(1-t_{2i})R_i \geq \tilde{V}$, $(i = H, L)$, where country type is known in period 2.[2]

[2]We assume throughout this analysis that the firm is maximizing profits in the location where they are earned, and thus cares only about the tax rates imposed where the firm is located. We therefore abstract from issues introduced by the possibility that the source country taxes foreign earnings when they are repatriated to the source country. For a discussion of the effects of the deferral provisions of U.S. tax law, see Thomas Horst (1977).

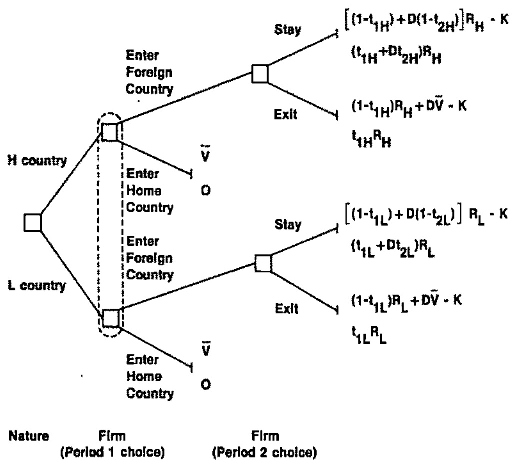


FIGURE 1

We follow Doyle and van Wijnbergen in assuming that in period 2, the country exploits the fixed investment made by the firm by raising the period-2 tax rate as high as possible without driving the firm out of the country. The period-2 tax rate will then make the firm indifferent between staying and leaving.[3] This maximum period-2 tax rate is denoted $\bar{t}_{2i}$, where

$$(1) \quad \bar{t}_{2i} = 1-(\tilde{V}/R_i), \qquad i = H, L.$$

Notice $\bar{t}_{2L} < \bar{t}_{2H}$ (since $R_L < R_H$) and $\bar{t}_{2L} < 1$ (since $\tilde{V} > 0$). We assume that $\bar{t}_{2L} > 0$, so that there exists a period-2 tax rate which will induce the firm to remain in the low-productivity country.

Given these second-period tax rates, the period-1 present value of the expected period-2 return for the firm does not depend upon the country's productivity and is given by $D\tilde{V}$, since any surplus above $\tilde{V}$ is extracted by either type of country. The firm will locate in a foreign country if the expected after-tax return from the investment at least equals the return from an investment in the home country, $\bar{V}$. Any period-1 tax rate chosen by a country must then satisfy the constraint

$$(2) \quad (1-t_{1i})R_i + D\tilde{V} - K \geq \bar{V}, \qquad i = H, L, P,$$

where i denotes the firm's belief about country type before entry. Equation (2) can be solved for the maximum period-1 tax rate, $\bar{t}_{1i}$. This rate makes the firm indifferent between entering a foreign country of type i and staying home:

$$(3) \quad \bar{t}_{1i} = (1-(K+\bar{V}-D\tilde{V})/R_i), \qquad i = H, L, P.$$

If the firm knows it is entering an H country, the highest tax rate the firm will be willing to pay is $\bar{t}_{1H}$. If the firm cannot discern the type of country it is entering, the highest tax rate the firm is willing to pay will be $\bar{t}_{1P} < \bar{t}_{1H}$.

The tax revenue collected by country i from the tax policy is $A_i = (t_{1i} + Dt_{2i})R_i$ $(i = H, L)$. We assume that host countries are revenue maximizers. In order for the country to be willing to offer a tax package, the present value of the tax revenue must be nonnegative. The lowest period-1 tax rate a country of type i is willing to offer is the one that makes tax revenue equal to zero, or

$$(4) \quad \hat{t}_{1i} = -D\bar{t}_{2i} = D((\tilde{V}/R_i)-1) \qquad i = H, L,$$

where the second equality follows from (1). The H country is willing to offer a larger subsidy than the L country $(\hat{t}_{1L} > \hat{t}_{1H})$ because an H country both charges a higher period-2 tax rate and collects higher period-2 revenue. Therefore, the H country can identify itself by offering a tax rate less than $\hat{t}_{1L}$. The firm knows that an L country could not offer such a large period-1 subsidy, because the second period return is too low to allow the L country to recover the subsidy.

[3]Similar results could be obtained in a model in which countries are able to commit to period-2 tax rates, and thus do not drive firms to indifference in period 2. Our working paper (1985) investigates the conditions under which commitment will be optimal in a model of technology choice.

II. Separating and Pooling Equilibria

If the H and L countries find it optimal to choose different tax rates, then the firm can infer the country's type from the tax rate it offers. An equilibrium in which this occurs is referred to as a separating equilibrium, since the two country types can be distinguished by their actions. If the H and L countries find it optimal to choose identical tax rates, then the observed tax rate reveals no information about the country's type. An equilibrium in which this occurs is referred to as a pooling equilibrium.

We have derived expressions for the maximum tax rates that the firm is willing to pay in a separating equilibrium ($\bar{t}_{1H}$) and a pooling equilibrium ($\bar{t}_{1P}$), and the minimum tax rate than an L country is willing to offer ($\hat{t}_{1L}$). In this section we examine the three types of equilibria that can result, depending on the relationships between these rates. The three cases are illustrated in Table 1.

If $\bar{t}_{1H} < \hat{t}_{1L}$, then the maximum period-1 tax rate that the firm is willing to pay in an H country is less than the minimum tax rate the L country will offer. The H country will then want to separate itself by offering the tax rate $\bar{t}_{1H}$ to the firm. Since this is the maximum tax rate the firm is willing to pay, the firm is driven to indifference between entering the H country and staying home, and the L country stays out. This is denoted as case A in Table 1.[4]

If $\bar{t}_{1H} > \hat{t}_{1L}$, then the H country cannot charge the tax rate the drives the firm's profits from entering to $\bar{V}$, because the L country would be willing to duplicate such an offer. If the L country makes such an offer, the firm cannot discern the country's type. The maximum tax rate the firm is willing to accept then drops to $\bar{t}_{1P}$. We will have two subcases to consider.

First, if $\hat{t}_{1L} > \bar{t}_{1P}$, then the L country is not willing to offer a tax rate low enough to induce the firm to enter. The H country can then offer a tax rate which identifies itself, but only if this tax rate is not set so high as to induce the L country to duplicate the offer. The firm is accordingly not driven to indifference by the H country in this case, because a tax rate above $\hat{t}_{1L}$ cannot be offered. Therefore, the firm benefits from the uncertainty about country type, because the country must give up some tax revenue (relative to the perfect information case) to keep an L country out of the bargaining. This is case B in Table 1.

Secondly, case C in Table 1 occurs if $\hat{t}_{1L} < \bar{t}_{1P}$. The L country now will offer tax rates low enough to induce entry. The L country and H country will then each offer the tax rate $\bar{t}_{1P}$ and the firm will be indifferent between entering and staying home. In this case the L country benefits from being confused with the H country. The L country gets more tax revenue (and the H country less revenue) than with perfect information about country types.

We now examine how the parameter values of the model will affect the likelihood of these three outcomes. A separating equilibrium with firms driven to indifference (case A) will occur if $\hat{t}_{1L} > \bar{t}_{1H}$. From (3) and (4) and using the fact that $\tilde{V}(1+D)+DK_d = \bar{V}$, we have

$$\hat{t}_{1L} > \bar{t}_{1H} \tag{5}$$

$$\Leftrightarrow \frac{D\tilde{V}}{R_L} + \frac{K+\tilde{V}+DK_d}{R_H} > (1+D).$$

This condition is more likely to be satisfied as the home-country options of the firm are more attractive (high $\tilde{V}$) and as the potential host-country options less attractive (R_L and R_H are low; K is high). The appearance of K_d in the numerator of (5) indicates that the composition of costs (between fixed and variable) in the source country also plays a role in the ability to separate. High levels of

[4] We assume that if the firm is indifferent between entering and staying at home, the firm chooses to enter the foreign country. Similarly, if the L country is indifferent between making an offer and not making an offer to the firm, it chooses not to make an offer. These outcomes could be ensured by perturbing the tax rates by small ε in the proper direction in order to break the indifference. The equilibrium tax rates can then be thought of as the limit of the perturbed tax rates as ε approaches zero.

TABLE 1

Case	H Country Strategy	L Country Strategy	Equilibrium	Outcome
A. $\hat{t}_{1L} > \bar{t}_{1H} > \bar{t}_{1P}$	$\bar{t}_{1H}$	–	Separating. Firm's profits in H country equal profits in source country.	Firm enters H. Firm does not enter L.
B. $\bar{t}_{1H} > \hat{t}_{1L} > \bar{t}_{1P}$	$\hat{t}_{1L}$	–	Separating. Firm's profits in H country exceed profits in source country.	Firm enters H. Firm does not enter L.
C. $\bar{t}_{1H} > \bar{t}_{1P} > \hat{t}_{1L}$	$\bar{t}_{1P}$	$\bar{t}_{1P}$	Pooling. Firm's expected profits abroad equal profits in source country.	Firm enters.

fixed costs (K_d) cause the firm's home-country option in period 2 to be low relative to the period-1 option (because a relatively high proportion of costs must be incurred in period 2). This makes it easier for an H country to separate itself through subsidies in period 1.

If (5) is not satisfied, then the equilibrium will be either a separating equilibrium with firms earning profits from entry or a pooling equilibrium. The separating equilibrium (case B) will occur if $\hat{t}_{1L} > \bar{t}_{1P}$. From (3) and (4) we have

$$(6)\quad \hat{t}_{1L} > \bar{t}_{1P}$$

$$\Leftrightarrow \frac{D\tilde{V}}{R_L} + \frac{K + \tilde{V} + DK_d}{R_P} > (1+D).$$

The same factors that make condition (5) likely to be satisfied (high values of $\tilde{V}$, K, and K_d; and low values of the R_i) will make (6) likely to be satisfied and hence a case B separating equilibrium more likely to appear. In addition, the case B separating equilibrium is more likely the lower is the probability of an H country (so that R_P is small). The pooling equilibrium is thus likely to appear when V, K, are small, the R_i are large, and p is large.[5]

III. Tax Holidays

We now examine the conditions under which the outcome of the tax negotiations will be a tax holiday for the firm. We consider two types of tax holidays. The first (weak) definition simply requires that the tax rate rise over time ($t_2 > t_1$), while the second (strict) requires the first period to involve a subsidy to the firm.

We first consider the case of perfect information about country type in order to provide a benchmark against which to compare the effects of uncertainty about country type. From (1) and (2),

$$(7)\qquad \bar{t}_{2H} - \bar{t}_{1H} = (DK_d + K)/R_H.$$

Condition (7) reveals a result attained by Doyle and van Wijnbergen. If fixed costs exist, the tax rate rises over time. This occurs because once the firm is in the country, the country exploits the fact that the firm has already incurred fixed costs. If fixed costs are equal to zero, (7) yields a tax rate that is constant over time, so that fixed costs are necessary for the appearance of a tax holiday with complete information. With respect

[5]A fourth case is also possible, in which $\bar{t}_{1H} < \hat{t}_{1H}$ and no foreign country can make an offer that a firm will accept. This case will occur if R_H is so low that the after-tax return in the home country exceeds the no-tax return in the host country.

to the strict definition of a tax holiday, (1) and (7) indicate that a subsidy will be obtained in the first period in the perfect information case only if $R_H - \tilde{V} - DK_d - K < 0$. Again, it is the presence of high fixed costs associated with the foreign investment that leads to the possibility of $\bar{t}_{1H} < 0$. Notice, however, that this stricter form of tax holiday is not ensured by the presence of fixed costs in the perfect information case.

In the presence of uncertainty, significantly different results appear. First, any separating equilibrium yields a strict tax holiday. As indicated by (5) and (6), the conditions that lead to a separating equilibrium depend on both the level of the fixed costs in the source country and the attractiveness of the host country relative to the source country. Therefore, it is possible to obtain a strict tax holiday even if fixed costs are zero, or $K = K_d = 0$. This will occur if the home country is sufficiently attractive relative to an H host country. This outcome is not possible in the case of perfect information, where fixed costs are necessary to obtain a weak tax holiday and relatively large fixed costs are required for a strict tax holiday.

In the case of a pooling equilibrium, the period-2 tax rate the firm pays will depend on whether the country turns out to be H or L. A comparison can be made between the first-period tax rate and the expected period-2 tax rate with pooling, $E(t_2) = p\bar{t}_{2H} + (1-p)\bar{t}_{2L}$. From (1) and (3),

$$(8) \quad E(t_2) - \bar{t}_{1P} = \frac{K + DK_d}{R_P} + \left(\frac{1}{R_P} - \frac{p}{R_H} - \frac{(1-p)}{R_L}\right).$$

The first term on the right side of (8) is identical to that obtained in the perfect information case. Since $(1/R)$ is convex in R, Jensen's inequality ensures that the second term in (8) must be negative. Therefore, a tax rate that rises over time (in the sense of expected tax rates) is less likely to occur in the presence of uncertainty about country type (given a pooling equilibrium) than in the perfect information case.

This analysis has assumed that the country will extract all the surplus from the firm in period 2. In addition, the firm receives no overall surplus in the case A separating equilibrium and the case C pooling equilibrium, since in both cases the firm is indifferent between entering and staying home. If several countries are introduced into the analysis, the countries may compete for the firm by offering lower tax rates. In this case, some form of a bargaining model in which firms and countries split the surplus from the foreign investment would be necessary.

IV. Conclusions

We have shown that tax holidays may be the outcome of a bargaining process between a firm and a country in which the country has better information about its productivity than the firm does. A high-productivity country may offer a tax holiday in order to identify itself. Low-productivity countries are unable to duplicate the tax holiday, because the future tax rate required to recoup the cost of the first-period subsidy would drive the firm out of the country when the productivity is learned. Unlike the complete information model of tax holidays, this result does not depend on the existence of fixed costs in order to generate a tax rate that rises over time. A tax holiday may occur in cases where fixed costs are zero if country type is unknown. Similarly, tax holidays in the incomplete information model may entail period-1 subsidies. This outcome may occur even without the high fixed costs required to generate a similar outcome in the complete information case.

REFERENCES

Bond, Eric W. and Guisinger, Stephen, "Investment Incentives as Tariff Substitutes: A Comprehensive Measure of Protection," *Review of Economics and Statistics*, February 1985, *67*, 91–97.

______ **and Samuelson, Larry,** "Bargaining with Commitment and the Theory of Tax

Holidays," Department of Economics Working Paper, Pennsylvania State University, 1985.

Caves, R., *Multinational Enterprise and Economic Analysis*, Cambridge: Cambridge University Press, 1983.

Doyle, Chris and van Wijnbergen, Sweder, "Taxation of Foreign Multinationals: A Sequential Bargaining Approach to Tax Holidays," Institute for International Economic Studies Seminar Paper No. 284, University of Stockholm, 1984.

Horst, Thomas, "American Taxation of Multinational Firms," *American Economic Review*, June 1977, *67*, 376–89.

Yuill, Douglas and Kevin Allen, *European Regional Incentive Programs 1981*, Glasgow: University of Strathclyde, 1981.

On the Profitability of Interruptible Supply

By GLENN LOURY AND TRACY R. LEWIS*

Recently much attention has been given to studying the optimal response of consumers who are threatened by intermittent interruptions in the supply of some valuable service or material. The literature on strategic stockpiling, for example, addresses the issue of how consuming countries can protect themselves against intentional or accidental interruptions in the supply of key mineral and petroleum products. Typically these analyses simply take as given the assumption that producers of a given product may at some time in the future stop supplies flowing to consumers. In this regard, it is important to know what incentives exist for a supplier to shut down and stop serving a market, assuming, of course, that the supplier can correctly anticipate the reaction of consumers to the threat of a supply embargo.

This paper examines the profitability of deliberate supply interruptions by a monopoly producer. The case for interrupting supply to increase profits relies on consumers storing the product in anticipation of eventual disruption of trade. Consumers may be induced to buy more of the product in early time periods, in order to maintain inventories as a hedge against future supply disruptions.

This increase in revenues resulting from inventory buildups must be traded off against the sales revenues lost during the time that supplies to consumers are cut off in order to determine the profitability of deliberate trade disruption. In what follows this tradeoff is investigated in the context of Loury's (1983) model of a stochastic on and off market. The model is described in Section I.[1] In Section II the profit function is derived for a monopolist who charges a constant price during periods when the market is on, and ceases all supply during periods when the market is off. We demonstrate that the result of random supply disruptions is, in effect, the same as if the monopolist was charging higher than normal prices during periods in which the market is not operating. From this analysis we are able to establish the following:

1) If the monopolist is unconstrained in the choice of the price he charges, then interrupting supply is never profitable.

2) When there are exogenous random supply disruptions, then under general conditions the monopolist will charge a price which is strictly lower than the single-period, profit-maximizing price during periods in which the market is operating.

3) A monopolist who is price constrained (say by regulation, or perhaps by the tariff policies of its trading partners) may find supply disruptions to be profitable.

I. The Model

The model we employ is that of Loury (1983). Assume that the market is described at any time as being in one of two possible states. When the market is "on," a commodity is supplied at some constant price p by the monopoly producer in each time period. For simplicity, zero production costs are assumed. Purchasers of the commodity may either consume it or store it at zero cost. When the market is "off," all supply of the commodity ceases, and the good is consumed by drawing down inventories which have previously been built up during periods when the market was operating.

*JFK School of Government, Harvard University, Cambridge, MA 02138, and Department of Economics, University of California, Berkeley, CA 94720, respectively. We thank Robin Cantor, Randy Curlee, George Sweeney, and Dave Trumble for valuable comments on an earlier draft.

[1] To ensure that this note is self-contained, we duplicate some of the analysis and results reported in Loury (1983) and C. Bergström, Loury, and M. Persson (1985). However, the reader is referred to these articles for a more complete description of the model.

It is assumed that the market switches between being on and being off according to a stationary Markov process. This means that the elapsed time, T_i, that the market is in state i (on or off) is exponentially distributed. In what follows we denote by h_1 the constant instantaneous rate of probability that the market switches from being on to being off, and by h_0 the instantaneous rate of probability that the market switches from off to on.[2] These switching probabilities, h_1 and h_0, reflect the reliability of service offered by the monopolist. In what follows, we assume that reliability is determined by factors (such as machine breakdowns, and interruptions in shipments of factors of production) that cannot be controlled by the monopolist.[3] In Section II, we have something to say about the incentives existing for suppliers to insure the reliability of their service to the market.

When the market is on, we denote by z_t the rate at which consumers purchase the commodity. The rate of consumption of the good is denoted by q_t, and the rate at which the current stock of inventories, denoted by S_t, changes is given by

$$\text{(1)} \qquad \dot{S}_t = z_t - q_t.$$

When the market is off, z_t is necessarily equal to zero.

It is assumed that the flow of instanteous utility to consumers denoted by $U(q,z)$ is given by

$$\text{(2)} \qquad U(q,z) = u(q) - pz,$$

where $u(\ )$ is increasing and strictly concave. Denote by $q(p) = u'^{-1}(p)$ the consumers' demand schedule for the product.

[2]Specifically, the probability that the market will remain in the on state until time T_1, given that it is on at time zero, is $h_1 e^{-h_1 T_1}$, and the probability that the market will remain off until time T_0, given that it starts that way is $h_0 e^{-h_0 T_0}$.

[3]An alternative interpretation of this model is that it is delays in delivery which cause the flow of supply to be interrupted periodically.

With these assumptions, Loury (1983) derives the following simple rules for optimal consumption and inventory accumulation for consumers subject to periodic disruption in trade. When the market is on, consumers choose a level of q equal to $q(p)$ so as to equate the marginal utility of consumption to the market price, p. In addition, at the moment the market begins operating, consumers instantaneously purchase inventories equal to S^* which they maintain throughout the period that the market is on.

When the market is off, consumers draw down their beginning of the period inventory, S^*, by choosing a time path for consumption so as to

$$\text{(3)} \qquad \underset{\{q(t)\}}{\text{maximize}} \int_0^\infty e^{-(r+h_0)}\left[u(q_t) - cq_t\right] dt$$

subject to

$$\int_0^\infty q_t\, dt \le S^*; \qquad c = \left(h_0/(r+h_0)\right)p,$$

where r is the rate of time discount. According to (3), when the market is off, consumers behave as though they have S^* units of a Hotelling-like exhaustible resource which they can extract at a constant unit cost c, and where the effective rate of time discount is given by $r + h_0$. The value for c represents the expected present value marginal cost of replenishing inventories once the embargo ends and the market resumes operation.

II. The Profitability of Supply Disruptions

The expected discounted profits earned by the monopolist depend on whether the market is currently on or off, and the existing inventories held by consumers. In what follows we employ the following notation: $\pi_1 \equiv$ expected present value of profits given that the market is currently on and $S = S^*$, where S is consumer inventory holdings; $\pi_0 \equiv$ expected present value of profits given that the market is currently off, and $S = S^*$; $\pi^* \equiv$ expected present value of profits given that the market is operating and $S = 0$; and

$\hat{q}(t) \equiv$ rate of consumption t units of time into an embargo.

We are interested in calculating π^* as we assume that initially the market is on, and consumers hold zero inventories. From the definitions given above it follows that

$$(4)\quad \pi^* = pS^* + \pi_1$$

$$(5)\quad \pi_1 = \int_0^\infty h_1 e^{-h_1 t} \times \left\{ \int_0^t e^{-rx} pq(p)\, dx + e^{-rt}\pi_0 \right\} dt$$

$$(6)\quad \pi_0 = \int_0^\infty h_0 e^{-h_0 t} \times \left\{ e^{-rt}\pi_1 + pe^{-rt}\int_0^t \hat{q}(x)\, dx \right\} dt.$$

Performing the indicated integration in (5) and (6) and simplifying one obtains

$$(5')\quad \pi_1 = \frac{pq(p)}{r+h_1} + \frac{h_1}{r+h_1}\pi_0$$

$$(6')\quad \pi_0 = h_0/(r+h_0) \times \left\{ \pi_1 + p\int_0^\infty \hat{q}(t)\, e^{-(r+h_0)t}\, dt \right\}.$$

One can solve for π_1 by substituting (6') into (5') to obtain

$$(7)\quad \pi_1 = \frac{r+h_0}{r+h_0+h_1}\,\frac{pq(p)}{r} + \frac{h_1}{r+h_0+h_1}\int_0^\infty h_0 e^{-(r+h_0)t}\frac{p\hat{q}(t)}{r}\, dt.$$

To calculate π^* in (4) we need to derive an expression for pS^*. Notice that

$$(8)\qquad pS^* = p\int_0^\infty \hat{q}(t)\, dt.$$

We can rewrite (8) as

$$(9)\quad pS^* = h_1/(r+h_0+h_1) \times \int_0^\infty \left[\frac{r(r+h_0+h_1)}{h_1} \right] \frac{p\hat{q}(t)}{r}\, dt = \frac{h_1}{r+h_0+h_1}\int_0^\infty (r+h_0) \times \left[\frac{r+h_1}{h_1} - \frac{h_0}{r+h_0} \right] \frac{p\hat{q}(t)}{r}\, dt,$$

which may be written as

$$(10)\quad pS^* = h_1/(r+h_0+h_1) \times \int_0^\infty (r+h_0)\, e^{-(r+h_0)t} \times \left[\frac{r+h_1}{h_1} - \frac{h_0}{r+h_0} \right] e^{(r+h_0)t}\frac{p\hat{q}(t)}{r}\, dt.$$

If we substitute (10) and (7) into (4) we obtain an expression for π^*:

$$(11)\quad \pi^* = \frac{r+h_0}{r+h_0+h_1}\,\frac{pq(p)}{r} + \frac{h_1}{r+h_0+h_1} \times \int_0^\infty (r+h_0)\, e^{-(r+h_0)t}\frac{p(t)\hat{q}(t)}{r}\, dt,$$

where

$$(11')\quad p(t) = \frac{ph_0}{r+h_0} + pe^{(r+h_0)}\left[\frac{r+h_1}{h_1} - \frac{h_0}{r+h_0} \right].$$

This expression for $p(t)$ turns out to satisfy $p(t) = u'(\hat{q}(t))$, that is, $p(t)$ is the implicit demand price along the optimal consumption path during an embargo. To verify this, first note that the solution to the consumers' maximization problem posed in (3) requires that

$$(12a)\qquad u'(\hat{q}(t)) - c = \lambda(t),$$

$$(12b)\qquad \dot{\lambda}(t) = (r+h_0)\lambda(t).$$

Combining (12a) and (12b) and substituting for $c = p\,h_0/(r+h_0)$ we obtain

$$(13)\quad u'(\hat{q}(t)) - (ph_0/(r+h_0)) = e^{(r+h_0)t}\left[u'(\hat{q}(0)) - \frac{ph_0}{r+h_0}\right],$$

or

$$(14)\quad u'(\hat{q}(t)) = p(h_0/(r+h_0)) + e^{(r+h_0)t}\left[u'(\hat{q}(0)) - \frac{ph_0}{r+h_0}\right].$$

Finally, notice that optimal inventory accumulation requires that

$$(15)\qquad p = u'(\hat{q}(0))h_1/(h_1+r).$$

To see this, suppose no embargo has yet begun, and one considers adding an additional unit to the (presumed optimal) stockpile now on hand. That unit could be consumed immediately upon the initiation of an embargo, generating at that time additional utility at the rate $u'(\hat{q}(0))$. Since the initial stock was presumed optimal, one would not have to replace this hypothetical additional unit during the next cycle of free trade. The expected present value of additional utility due to adding this unit, at the time it is added equals $u'(\hat{q}(0))h_1/(r+h_1)$. Unless this term is equal to the price of an additional unit of the resource, the *size* of the stockpile could not have been optimal, as supposed.[4]

Substituting (15) into (14) and combining this with (11′) yields

$$(16)\quad p(t) = u'(\hat{q}(t)) = ph_0/(r+h_0) + e^{(r+h_0)t}\left[\frac{p(r+h_1)}{h_1} - \frac{ph_0}{r+h_0}\right],$$

which proves our earlier assertion that $p(t)$ is the implicit demand price along the optimal consumption path during an embargo. Notice that since $u'(q(p)) \equiv p$, it follows that $\hat{q}(t) = q(p(t))$, so we may rewrite the expression for π^* in (11) as

$$(17)\quad \pi^* = \left(\frac{r+h_0}{r+h_0+h_1}\right)\frac{pq(p)}{r} + \left(\frac{h_1}{r+h_0+h_1}\right)\int_0^\infty (r+h_0)\,e^{-(r+h_0)t} \times \frac{p(t)q(p(t))}{r}\,dt.$$

The first term on the right-hand side of (17) represents the expected discounted value of revenues earned while the market is operating. The second term measures the expected present value of revenues earned due to consumers' purchase of inventories to draw upon while the market is not operating. Notice that the second term is a weighted sum of the "implicit" market revenues, $p(t)q(p(t))$, generated during the embargo period.

Several conclusions follow immediately from inspecting (17). First, let $\bar{p}$ be the price (assuming it exists) which maximizes the flow of revenues $pq(p)$. Then we have

$$(18)\ \pi^* = \frac{r+h_0}{r+h_0+h_1}\frac{pq(p)}{r} + \frac{h_1}{r+h_0+h_1}\int_0^\infty (r+h_0)\,e^{-(r+h_0)t} \times \frac{p(t)q(p(t))}{r}\,dt$$

$$\le \frac{r+h_0}{r+h_0+h_1}\frac{\bar{p}q(\bar{p})}{r} + \frac{h_1}{r+h_0+h_1} \times \int_0^\infty (r+h_0)\,e^{-(r+h_0)t}\frac{\bar{p}q(\bar{p})}{r}\,dt$$

$$= \bar{p}q(\bar{p})/r.$$

[4] The implications of this are explored further in Bergström et al.

Equation (18) holds with strict inequality provided $q(p)$ is not unitary elastic everywhere, or equivalently that $u(q) \neq a + b \log q$. Thus according to (18), supply disruptions are generally unprofitable for a monopoly supplier who is unconstrained in his choice of price during periods in which the market is operating.[5]

Second, if we assume that total revenue, $pq(p)$ is strictly concave in price, then we can show that the price which maximizes π^* is strictly less than $\bar{p}$, the price which a conventional monopolist would charge in the absence of supply interruptions. Let p° maximize π^*. A necessary condition for maximization is that

$$(19)\quad 0 = \frac{d\pi^*}{dp} = \frac{r+h_0}{r+h_0+h_1}\frac{MR(p^\circ)}{r} + \frac{h_1}{r+h_0+h_1}\int_0^\infty (r+h_0)\,e^{-(r+h_0)t} \times \frac{MR(p(t))}{r}\frac{dp(t)}{dp}\,dt,$$

where $MR(p)$ is marginal revenue expressed as a function of price and

$$dp(t)/dp = h_0/(r+h_0) + e^{(r+h_0)t}\left[\frac{r+h_1}{h_1} - \frac{h_0}{r+h_0}\right] > 0$$

from equation (16). Equation (16) also implies that $p(t) > p$ for all t. Given our assumption that total revenue is strictly concave, it follows that

$$MR(p) < MR(\bar{p}) = 0 \quad \text{for } p > \bar{p}.$$

This implies that $p^\circ < \bar{p}$ in order to satisfy (19). Otherwise with $p^\circ \geq \bar{p}$ we would have $MR(p^\circ) \leq 0$ and $MR(p(t)) < 0$ (since $p(t) > p^\circ \geq \bar{p}$) thus implying a violation of (19). Hence market prices will be set below the instantaneous revenue-maximizing price whenever the market is subject to supply disruptions. The lower price encourages consumers to increase their purchase of inventory stocks by an amount sufficient to offset its negative effect on revenue flow when the market is on.

Finally, and perhaps most important, is the observation already noted above, that $p(t) > p$, according to equation (16). Recalling the expression for π^* in equation (17), this means that the presence of recurrent supply interruptions has the same effect as if the monopolist were to charge consumers higher prices than p during periods when the market is off. This is important for a producer who is constrained to choose prices in the inelastic region of the demand curve because of regulations, or perhaps because of tariff policies which effectively limit the price which may be charged. If the ceiling price constraining the monopolist is sufficiently low, it seems clear from equation (17) that profits may increase as a result of supply disruptions. For example, consider the special case where demand is inelastic and of constant elasticity $\eta < 1$, and where the monopolist is constrained to charge a price $p \leq p^*$, where p^* maximizes Π^* as in equation (19). In the absence of supply disruptions, the monopolist would always charge the ceiling price $p = p^*$, since profits are strictly increasing in p. This would yield present-value discounted profits equal to $p^*q(p^*)/r$. With supply disruption, the monopolist would also set $p = p^*$ (whenever the market operates) given the definition p^*. However, in this instance, (16) and (17) would imply that $\pi^* > p^*q(p^*)/r$, since the "effective" price during periods when the market is not operating would strictly exceed p^* and profits are strictly increasing in price. We note that the possibility that supply disruptions may result in increased profits for a regulated monopolist is easily illustrated by numerical examples for more general specifications of demand.

[5]S. R. Aiyagari and R. G. Riezman (1985) also demonstrate the unprofitability of supply disruptions in a two-period example of a linear quadratic model. Our result corrects an error in Loury (1983), whose equation (14) and subsequent discussion is incorrect.

III. Conclusion

The import of our analysis is to suggest a rationale for certain observed cases where the supply of some good or service is sometimes temporarily interrupted. One implication of our results is that disruptions in supply may be profitable for a monopolist in cases where consumers can store inventories at low cost in anticipation of supply shortages, and the monopolist is constrained to operate over the inelastic portion of the demand schedule when the market is operating. In environments where these conditions do not hold, we would expect monopoly producers to undertake whatever measures are available to insure against interruptions in supplies to the market. We have employed a rather special model to derive our results in which we assume the market vacillates randomly between being on or off.

While we have not worked out the details, it seems clear to us that our results about interruptable markets would also apply in situations where a monopolist can control and vary his supply to the market over time. For example, imagine that a monopolist is able to commit itself to a time path of prices that it will charge consumers now and in the future. In this case, it seems obvious that if the monopolist is unconstrained in his choice of prices, he will agree to supply the market at the single-period, profit-maximizing price forever. However, a price-constrained monopolist may find it profitable to reduce (or totally eliminate) supplies to the market in future time periods as a device whereby he can effectively raise the price that he charges the market during times when supplies are rationed or completely cut off. This of course requires that the monopolist can credibly commit to reducing future supplies in some verifiable way. An interesting issue which we do not explore is, given that the monopolist determines to employ strategies of periodic interruption of sales, what combination of frequency and duration of interruptions is profit maximizing? (That is, which choice of h_0 and h_1 maximize Π^*?)

REFERENCES

Aiyagari, S. R. and Riezman, R. G., "Embargoes and Supply Shocks in a Market with a Dominant Seller," in T. Sargent, ed., *Energy, Foresight and Strategy*, Baltimore: Johns Hopkins University Press, 1985.

Bergström, C, Loury, G. and Persson, M., "Embargo Threats and the Management of Emergency Reserves," *Journal of Political Economy*, February 1985, *93*, 26–42.

Loury, G., "The Welfare Effects of Intermittent Interruptions of Trade," *American Economic Review Proceedings*, May 1983, *73*, 272–82.

Final Voting in Legislatures

By W. MARK CRAIN, DONALD R. LEAVENS, AND ROBERT D. TOLLISON*

In representative democracies, such as the United States, legislatures provide the transmission mechanism through which pressure from private interests becomes public policy. Considerable attention has been given in the literature to explanations of the relevant forces that appear to be driving the legislative process. For example, much research has focused on the relative impact of economic vs. ideological influences on congressional voting behavior. In this approach, the way that legislators vote on proposed legislation is modeled as a function of the preferences of various economic and ideological interests groups, including the legislator's own preferences for wealth and ideology (James Kau and Paul Rubin, 1979; Joseph Kalt and Mark Zupan, 1984; Sam Peltzman, 1985).

Missing from this approach is the idea that when legislatures are the transmission mechanism, they are costly and imperfect organizations for generating political influence (Gary Becker, 1983). As such, rules and institutions will emerge that are related to problems of internal control within the organization of a legislature. In this paper, we focus on the role of floor voting from the standpoint of legislator organization and control. We seek to expand the interpretation of the meaning of floor voting activity by examining the timing, sequence, and outcomes of such votes. Specifically, we look at final floor voting in the U.S. Congress. The patterns described in the analysis below suggest that a broader analytical perspective on the economic function of floor voting is required. The findings also suggest that to identify more precisely the forces that are driving legislator voting behavior, it is important to recognize the role of legislative transactional costs and institutional constraints.

In Section I, the conceptual framework for the empirical results is discussed in more detail. The purpose is not to develop a full-blown theory of legislative organization, rather, it is to focus the reader's attention on several hypotheses about the function of final floor voting as a device for controlling legislator behavior within the legislature. Empirical results, including an explanation of the timing and sequence of final votes on bills, are reported in Section II. The data for these tests are drawn from legislative activities in the U.S. House of Representatives during the 96th and 98th Congresses. Some concluding remarks are offered in Section III.

I. Voting and Legislative Organization: Some Economic Principles

Our purpose in this section is simply to prepare the groundwork for the empirical analysis presented in the following section. This discussion is in no sense a formal model of legislative behavior. It is rather an application of some principles of economic organization to derive some empirical inferences about legislative voting patterns. We suggest several explanations for the role of floor voting, other than the usual view that voting represents the final tallying of legislator preferences on public policy.

We begin the discussion with the concept that the nature of collective decision making, per se, makes the familiar principles of team production applicable to legislative behavior (Armen Alchian and Harold Demsetz, 1972). By virtue of the requirement of a majority vote to pass legislation, the actions of individual legislators are linked, and the usual implications of team production follow (Crain and Tollison, 1980, 1982; Arleen Leibowitz and Tollison, 1980). In other words, shirking by individual legislators will

*Center for Study of Public Choice, George Mason University, Fairfax, VA 22030. We are grateful to the referees and to Robert Haveman for useful comments. The usual caveat applies.

be a problem, and the monitoring and management of legislator behavior will be an important aspect of the legislature and its organization. We can illustrate the team production argument with respect to how bills are selected for final approval by the legislature.

Final voting on legislation could take place at any time during the legislative session, from the first day on. But the principles of team production imply that final voting will occur primarily at the *end* of the session. This follows because of the potential for legislator shirking. Deals are negotiated during the legislative session among legislators, the leadership, and interest groups. Once it is clear that these "markets" have cleared, the leadership must arrange final votes to consummate each deal. Deals could be voted upon as soon as it was clear what was wanted, but in a team production setting this leads to the potential for reneging. A simple way to think about the problem of reneging is that once a legislator gets his bill passed, he goes home without sticking around to carry out the quid pro quos that led to the passage of his bill. Alternative ways to renege are simply to change one's vote on subsequent bills, or not to work as hard for the passage of other bills. Hence, the first testable implication is that most final votes will be taken at the end of the legislative session in marathon, widely publicized meetings, where the opportunities for reneging are low.

The legislature will be a more effective organization the more closely rewards are tailored to individual legislator productivity. The latter can take a variety of forms, but for the purpose of analyzing final votes, individual legislator productivity is defined as the propensity of a legislator to keep his political bargains. Those legislators who can be counted on to keep their bargains receive differentially higher rewards from the leadership. Solutions to the problem of differential rewards for differential productivity can be found in many areas of economic theory. For example, an hypothesis from the screening literature is that legislators, like employees in certain types of job settings, are screened for quality by the employers within the organization. In the legislative context, information about the qualities of party loyalty and trustworthiness is important to the party leaders. Institutions like the committee system and seniority rules are examples of loyalty filters in legislative organizations (George Akerlof, 1983). The party leadership can influence voting by using votes as a criterion for making superior committee assignments.[1] There are two implications that follow from applying the filtering framework: (a) in the scheduling of final votes, bills sponsored by more trustworthy legislators will stand higher in the floor voting queue, and (b) these same legislators will have a higher probability of obtaining the passage of their bills in committee and floor voting.

We proxy trustworthiness by measures of legislator seniority and influence. The members of the legislature who exhibit more loyalty to the leadership have their bills passed first because the leadership knows from experience that they can be relied upon to carry through on the bargains they made with other legislators. In other words, earlier and more certain passage of bills is a premium paid to those legislators who have passed through a party loyalty filter. Their differential reward is a higher expected present value for the laws they sponsor (they are more likely to be passed and to be passed sooner in final voting), with the accompanying implication of better reelection prospects and so forth.[2]

There are other competing explanations for why the pattern of voting on bills will be influenced by political considerations that stem from the internal organization of the legislature. A theme in political science literature is that more politically powerful legislators get their bills passed sooner and more

[1] See Charles Bullock (1985) and the references therein for a summary of the political science literature on this thesis. These studies examine congressional committee assignments as a function of past voting behavior.

[2] A more complete theory of bill passage would include a framework to explain the date of introduction as well as passage. We make a preliminary effort to come to grips with this problem empirically in the next section by using the loyalty/legislative capital model to explain a bill's success at the stage of committee voting.

easily. The basis of their power may not be seniority but, for example, a charismatic personality, or the fact that they represent an influential constituency. The fact, however, that power and seniority have a natural affinity means that a political power theory carries many of the same implications as the above discussion for final voting on bills.[3] The intricacies and sources of power are driven in this type of framework by the concept of accumulating political capital. Since expertness and legislative capital are built over time, legislative competence will also naturally be related to seniority. More senior legislators have been able to acquire larger stocks of legislative capital, and hence we would expect them to get their bills passed sooner. Thus, a legislative capital framework also leads to an implication that floor voting activity depends on elements related to legislative organization.[4]

There is one final approach that we can suggest. It is the idea that voting on the most important bills will occur at the end of the legislative term, so that legislators can squeeze the highest return from these transactions. In this view, the passage of legislation is analogous to the holdout problem —the longer you wait, the more rents you can squeeze. This bargaining approach also predicts that final votes will be at the end of the session. There is, however, a way to distinguish this hypothesis from the team production and loyalty filtering hypotheses. The bargaining power theory suggests that returns are greatest by waiting the longest, and if this is so, the bills of the most senior and powerful legislators should be voted on last. The earlier arguments, in contrast, suggest that the control of individual legislator incentives is paramount to producing final bills, and that for this reason the bills of the most senior legislators are voted on first in the queue of final votes.

There are undoubtedly other examples in this spirit. Our point is to stress the importance of political institutions and their possible effects on floor voting behavior. The key to understanding and interpreting the information embodied in voting outcomes requires more focus on these organizational problems.

II. Empirical Analysis: Voting on U.S. House Bills

A. *The Timing and Sequence of Final Votes*

The empirical analysis of final floor voting is based on data on the U.S. House of Representatives. First, we examine the timing of final voting on the House floor. In this case the data are from U.S. Congressional Research Service (1981).

The 96th Congress commenced on January 15, 1979, and adjourned on December 16, 1980. A total of 8,455 House bills were introduced in the 96th Congress. Of these, 6,232 (74 percent) were introduced in the first session and 2,223 (26 percent) in the second session. Most bills were introduced early, and were around to be debated and voted on at an early stage. Yet most voting took place at the end of the available legislative work time. Of the 8,455 bills introduced, 560 (6.7 percent) made it through the chamber and passed on the House floor. Of the 560 bills passed, 250 (45 percent) were passed in the first session and 310 (55 percent) in the second session.

The 96th Congress worked a total of 322 days, excluding holidays and other days not in session such as Fridays and weekends, so a bill that was passed on the first day would have had 322 legislative days remaining. As a crude conversion formula, a legislative work day was equivalent to about 2.3 calendar days in this Congress.

The median for all bills passed by the House in the 96th Congress was legislative

[3] See Richard Fenno (1973), Morris Fiorina (1974), and John Ferejohn (1974) for contributions to this literature.

[4] The legislative capital argument embodies a simultaneity problem as do the other frameworks outlined. That is, effective legislators introduce good bills, good bills pass, passage of one's bills increases reelection prospects, reelection means more seniority, and so on. Our approach in the empirical section is somewhat more limited in scope than this. For a given Congress, we seek to explain the order of final votes by legislator characteristics. We recognize, in general, however, that the simultaneity issue is not trivial.

work day 207, with 116 legislative days remaining. On either side of this day, 280 bills were passed. Thus, roughly one-half of the successful bills were passed in the final one-third of the working days. The mean number of days remaining for all bills passed is 129, with a standard deviation of 85 days. Although this mean is not significantly different from the end of the legislative term (zero days remaining), it is significantly different from the beginning of the term (322 days remaining).

This overview of the raw data suggests that the timing of final votes is skewed towards the end of the legislative term. However, there is considerable variation in the timing of final votes. Not all votes occur at the end, nor would this pattern be possible due to a time constraint. Some bills have to be voted on before others.

The burden of our effort to apply some economic principles to legislative organization is to explain the order of the final votes that we observe. For this purpose we use the following model:

(1) *LEGISLATIVE DAYS REMAINING AFTER A BILL PASSES*
= f[SPONSOR CHARACTERISTICS, NUMBER OF COMMITTEES THAT CONSIDERED A BILL, LEGISLATIVE DAY BILL INTRODUCED].

The dependent variable, *LEGISLATIVE DAYS REMAINING*, measures how close to the end of a legislative term a bill passes. As *LEGISLATIVE DAYS REMAINING* declines, a bill is passed later in the Congress.

The *SPONSOR CHARACTERISTICS* variable is central to our analysis. Several of the hypotheses discussed above suggest that bills sponsored by more senior and well-positioned legislators will be passed earlier in the term. In the filtering approach, earlier passage is expected because it is less likely that more trustworthy members will stop being productive after their deals are consummated. In the team production approach, earlier passage is a greater reward to more productive members because the present value of a bill is higher the earlier it passes. By these same two arguments, bills sponsored by members whose seniority and committee positions are low will be passed later in order to make sure that the work gets done before rewards are handed out. In contrast, the holdout hypothesis predicts that bills sponsored by the more powerful members will be passed later. We proxy *SPONSOR CHARACTERISTICS* in several ways, including simple legislator tenure, party affiliation, subcommittee chairmanship, and an index of legislator seniority based on the importance of committee assignments and rank within those committee assignments.[5]

The *NUMBER OF COMMITTEES* variable controls for the problem of overlapping jurisdictions. A bill that must be considered in more than one committee will naturally reach the stage of final voting later. The *DAY INTRODUCED* variable functions as a time index for the analysis. Obviously, a bill must be introduced before it can be voted on, and bills that are introduced late in a session cannot be voted on until late.

In Table 1, the model is estimated on the 560 bills passed by the House in the 96th Congress.[6] We present results for five specifications of the model. Keep in mind that as a bill is passed closer to the end of

[5]The seniority index for an individual legislator is computed as follows:

$$\sum_{j=i}^{n}\{1-[(RANK_j-1)/NOFMEMB_j]\}\times[3-(2MINOR_j/MAJOR_j)],$$

where $RANK_j$ = the sponsor's rank on committee j, $NOFMEMB_j$ = the number of members on committee j, $MINOR_j$ = the number of minority members on committee j, $MAJOR_j$ = the number of majority members on committee j, and n = number of committee assignments for the sponsor. See Crain and Tollison (1981) and Leavens (1984) for other applications of this approach to assessing the influence of legislator seniority.

[6]In our approach, a bill is a bill, when surely not all bills are created equally. Bill content is an important issue, but one that is beyond the scope of this paper to address.

TABLE 1—*OLS* REGRESSION RESULTS FOR LEGISLATIVE DAYS REMAINING AFTER A BILL PASSES: ALL BILLS PASSED

Independent Variables	Coefficients[a] Model A	Model B	Model C	Model D	Model E
Constant	213	207	212	209	213
	(23)	(22)	(21)	(20)	(21)
LEGISLATIVE DAY	−.71	−.71	−.71	−.71	−.71
BILL INTRODUCED	(28)	(28)	(27)	(27)	(27)
NUMBER OF COMMITTEES	−11.08	−10.84	−9.73	−9.89	−9.68
THAT CONSIDERED BILL	(1.80)	(1.76)	(1.56)	(1.60)	(1.56)
SPONSOR'S TENURE	2.02	1.57	–	–	–
	(3.73)	(2.72)			
SPONSOR A MEMBER OF	−9.79	−1.9	–	−3.40	–
MAJORITY PARTY	(1.19)	(.22)		(.38)	
SPONSOR A SUBCOMMITTEE	–	13.10	–	16.37	–
CHAIRMAN		(2.24)		(2.80)	
SPONSOR'S SENIORITY INDEX	–	–	5.57	2.87	−.18
			(2.41)	(1.17)	(.04)
Interaction of *SPONSOR'S SENIORITY INDEX* and *MAJORITY PARTY*	–	–	–	–	6.02
					(1.72)
Adjusted R Square	.58	.59	.58	.58	.58
$F_{(\text{d.f.})}$	198	161	255	158	193
	(4,555)	(5,554)	(3,556)	(5,554)	(4,555)

Source: U.S. Congressional Research Service (1981).
[a]The t-statistics are shown in parentheses.

the session, the dependent variable gets smaller; a larger value means earlier passage. The *DAY INTRODUCED* variable appears in the expected direction and is strongly significant. The later a bill is introduced, the later it is passed. The coefficient values are calibrated in days. If a bill is introduced 10 legislative days earlier, its date of passage will be moved ahead 7 days. The gain is not day-for-day.[7] The *NUMBER OF COMMITTEES* variable appears in the predicted direction, and approaches a reasonable level of statistical significance in all models. Each additional committee to which a bill is referred delays final passage by about 10 days (roughly three weeks of calendar time).

The variables employed to proxy sponsor loyalty and legislative capital appear in the expected direction. Simple legislator tenure (*SPONSOR'S TENURE*), included in Models A and B, has a positive and highly significant sign. The bills sponsored by more senior legislators are passed sooner. The estimates indicate that for every year of past legislative service, a legislator's bill is passed 2 days sooner (see the coefficients in Models A and B in Table 1). Thus, a five-term congressman would have his bills passed 20 working days sooner than a freshman legislator, on average.

Party affiliation of the sponsor does not seem to matter at the stage of final floor voting. Indeed, Model A suggests that minority sponsors obtain earlier passage. This result is an artifact of the legislative process. Few minority bills make it beyond committee for a final vote, and those that do obviously have wide bipartisan support. In the 96th Congress, 47 bills sponsored by minority party House members were cleared by committees for a final vote. Out of these, 46 passed on the House floor.

The variable *SUBCOMMITTEE CHAIRMAN* exerts a powerful influence over the

[7]This variable is obviously a highly significant explainer of the date of bill passage. Yet the seniority variables, as reviewed below, also perform well and have pronounced marginal impacts on the timing of final passage.

timing of passage. The estimates presented in Table 1 (Models B and D) suggest that subcommittee chairmen have their bills passed on average about 14 legislative days sooner than their colleagues. This finding is consistent with several arguments. Committee chairmanships are typically awarded to members who have established the best reputations for party loyalty and who have accumulated considerable legislative capital as well.

Finally, the coefficients on the *SPONSOR'S SENIORITY INDEX* (Models D and E) suggest that relative committee rank and committee assignments are important determinants of the earlier passage of legislation. In Model C, when this variable is included without the other proxies for *SPONSOR CHARACTERISTICS*, it is positive and highly significant. The magnitude of the effect of the seniority index is more difficult to summarize since the coefficient can reflect the importance of different committee assignments as well as different seniority rankings on a given committee.[8]

In Model E we enter the *SENIORITY INDEX* for majority party members separately from minority party members. In this case we find that there is a difference between the two parties. For minority party members, there does not appear to be an ordering of final passage dates based on sponsor seniority. However, for majority party members, the *SENIORITY INDEX* retains its statistical significance, and the magnitude of the coefficient is slightly bigger than the estimated coefficient in Model C.

Overall, the models in Table 1 are highly significant, and explain about three-fifths of the variation in the timing of passage of bills.[9] It should be stressed before proceeding that the bargaining power, or holdout theory, does not fare well in these results; more legislator influence leads to earlier, not later, passage of a bill.

B. *Predicting Outcomes on Final Votes*

Not all bills that clear committees pass in final voting. In the 96th Congress, 597 House bills were reported from committee for consideration on the floor. Only 13 of these bills (2 percent) failed to muster a majority vote. However, 89 bills (15 percent) were never brought to a floor vote by the leadership.

These data provide a basis for another application of the economic hypotheses discussed above. Since not all the bills that reach the floor pass or get voted on, the hypotheses should be useful for analyzing the characteristics of the bills that are less likely to pass. For example, when a bill does not pass or is tabled, it can be thought of as resulting from a contractual breakdown of some sort in the legislative marketplace. The sponsor and other supporters of a bill may have engaged in reneging or other noncooperative behavior in carrying through on their commitments. The leadership needs to limit such behavior, and one way to discipline reneging is to refuse to call up a member's bill for a final vote or not to support such bills when the floor vote is held. In other words, when shirking takes place, payoffs are withheld by not passing a member's bill. Our approach should be able to *predict* which legislator's bills will suffer this fate.

Our experiment is the following. We took Model A from Table 1, and reestimated the coefficients using discriminant analysis.[10]

[8]For example, the tenth ranking member on the Energy and Commerce Committee could expect his bill to be passed about 3 working days sooner than the tenth ranking member of the Post Office and Civil Service Committee. Moreover, bills sponsored by the tenth ranking member on Energy and Commerce will be passed more than 4 days earlier than the lowest ranking member (twenty-seventh) on this committee.

[9]We replicated the results in Table 1 on a slightly different data base. We included in the five regressions only bills that were reported from committees for floor action. Out of the 560 bills that passed, 65 were not reported from a committee. Using this set of 495 bills, the results are unchanged. In particular, *SPONSOR'S TENURE* continues to perform as expected, as do the *SUBCOMMITTEE CHAIRMAN* and the *SENIORITY INDEX* variables.

[10]We used Model A for this purpose because it was the simplest approach in Table 1. It, for example, has the most straightforward definition of legislator tenure. The classification results using Models B–E are highly similar to those obtained using Model A, and do not

TABLE 2—CLASSIFICATION RESULTS ON PREDICTED FLOOR OUTCOMES

Actual Floor Outcome	No. of Cases	Predicted Floor Outcome[a]		
Class 1		Not Considered	Rejected	Passed
Not Considered	89	48 (54)	20 (22)	21 (24)
Rejected	13	5 (39)	2 (15)	6 (46)
Passed	495	179 (36)	98 (20)	218 (44)
Percent of Predicted Outcomes Correctly Classified = 45 percent				
Class 2		Not Considered		Considered
Not Considered	89	54 (61)		35 (39)
Considered	508	196 (39)		312 (61)
Percent of Predicted Outcomes Correctly Classified = 61 percent				
Class 3		Not Passed		Passed
Not Passed	102	61 (60)		41 (40)
Passed	495	187 (38)		308 (62)
Percent of Predicted Outcomes Correctly Classified = 62 percent.				

Source: See Table 1.
[a]Percentage correct of cases predicted is shown in parentheses.

Three ways of classifying the dependent variable were examined, and the results are summarized in Table 2. Class 1 divides outcomes in terms of whether bills were passed, rejected, or not considered; Class 2 divides outcomes in terms of whether bills were considered or not considered; and Class 3 divides outcomes in terms of whether a bill passed or did not pass. This experiment allows us to evaluate the potential of the model to explain the outcomes of final floor voting.[11]

We adopt a classification strategy in Table 2 that is naive and biased against the model. We assume that the prior probability that a bill will land in any of the selected categories is the same. This is obviously not so, and we know that we could achieve stronger results by using additional information to weight prior probabilities, for example, the proportion of bills that passed in recent Congresses. The naive approach reported in Table 2 is sufficient, however, to illustrate the potential of this analytical framework to classify correctly the fate of bills in floor votes.

In Class 1, where the categories are tripartite, the model correctly identifies 54 percent of the bills that were not considered. In Class 2, the model identifies 61 percent of the bills that were not considered, where a binary classification is used to divide bills into those that were considered (passed or

change the results in Table 2 in any meaningful way. For example, comparing Models B–E to the results for Model A, the percent of predicted outcomes correctly classified is, respectively, 61, 61, 60, and 61 percent. Moreover, we obtain slightly better results in our ability to identify those bills that were not passed using Models C and E. Model C correctly identifies 63 of the 102 bills not passed, and Model E identifies 64 of these bills. Models B and D correctly identify 61 and 62 of the 102 bills not passed. We would be happy to provide these results on request.

[11]The results of the discriminant analysis using Model A can be summarized as follows. The F-significance level for each of the three classifications was 1 percent, and the Wilks *lambda* statistics were well within the acceptable range. These results and the standardized coefficients and the univariate F-statistics for the individual coefficients are available on request.

TABLE 3—CLASSIFICATION RESULTS ON PREDICTED COMMITTEE OUTCOMES

		Predicted Committee Outcome[a]	
Actual Committee Outcome	No. of Cases	Not Passed	Passed
Not Passed in Committee	1721	1026 (60)	695 (40)
Passed in Committee	162	56 (35)	106 (65)
Percent of Predicted Outcomes Correctly Classified = 60%			

Source: Legi-Slate (1985).
[a]See Table 2.

failed) versus those not considered (tabled). In Class 3, the division is between bills that passed versus bills that did not pass (failed or not considered). We are able to identify bills in the not-passed category 60 percent of the time. This means that our model correctly identifies 61 of the 102 bills that did not pass. By comparison, a random model (50–50) would predict that 298 of the bills would fail.

Why is this an important result? It is important because we are dealing with the exception rather than the rule. Clearly, once reported from committee, most bills pass. But the 17 percent that did not pass puts an explanatory burden on the economic arguments. That is, can the alternative hypotheses derived from economic theory be applied to predict which bills will fall into this rare category of legislative behavior? (Yes.)

C. *Predicting Outcomes in Committees*

The hypotheses discussed in Section I also suggest a way to identify which bills will succeed through the committee process. Essentially, the same factors that explain the organization of final voting should explain the fate of bills as they are introduced and referred to committees. More senior and more trustworthy legislators' bills will be more likely to clear the committee stage for final votes on the House floor. Just as earlier passage is a way to reward legislator loyalty and performance, so too is favorable treatment of a legislator's proposals by the committee system.

To examine this implication and to provide more generalized evidence on our approach, we selected a random sample of bills introduced in the 98th Congress. The source for the second data set is Legi-Slate (1985). There were 6,444 House bills introduced in this Congress, out of which we drew a random sample of 2000 bills. We excluded 107 private bills. Thus, our final sample consists of 1,893 bills that were referred to House committees.

Our procedure with this second data set is straightforward. Again, we take Model A from Table 1, update the right-hand side variables to the 98th Congress, and recast the dependent variable in discrete terms, in this case, in terms of whether a bill passes or fails at the committee stage.[12] A bill can be voted down at the committee stage or simply never be acted upon.

The ability of the model to classify bills according to passed vs. not passed in committee is summarized in Table 3. We do not have to dwell upon the results. The legislator loyalty/legislator capital arguments offer statistically useful approaches to the problem of identifying which bills survive committees. To illustrate, there are 162 bills (8 percent) in the sample that actually passed committee. Model A correctly identifies 106 (65 percent) of those bills, taking again the naive a priori view that every bill in the

[12]We use Model A from Table 1 for the reasons cited in fn. 11 above. The results of the discriminant analysis for classification of pass/not passed at the committee stage were: the F-significance level was .0001, and the Wilks *lambda* was .968. These results, the univariate F-statistics, and individual coefficients are available on request.

sample has an even chance of passing committee.

III. Concluding Remarks

The legislature is an important institution in a mixed capitalist economic system, and understanding and predicting its behavior is a central task for economic theory. Our purpose in this paper has been to suggest several alternative ways to proceed in this task. The findings for the U.S. House of Representatives suggest that the timing, sequence, and outcomes of final floor votes are influenced by organizational considerations. While we do not claim to have pinpointed an exact explanation for final voting behavior in legislatures, we think we have provided enough evidence to broaden the debate about how to interpret its meaning.

REFERENCES

Akerlof, George A., "Loyalty Filters," *American Economic Review*, March 1983, *73*, 54–65.

Alchian, Armen A. and Demsetz, Harold, "Production, Information Costs, and Economic Organization," *American Economic Review*, December 1972, *62*, 777–95.

Becker, Gary, "A Theory of Competition Among Pressure Groups For Political Influence," *Quarterly Journal of Economics*, August 1983, *98*, 371–400.

Bullock, Charles S. III, "U.S. Senate Committee Assignments: Preferences, Motivations, and Success," *American Journal of Political Science*, November 1985, *29*, 789–808.

Crain, W. Mark, and Tollison, Robert D., "The Sizes of Majorities," *Southern Economic Journal*, January 1980, *46*, 726–34.

______ **and** ______, "Representation and Influence: Reply," *Journal of Legal Studies*, January 1981, *10*, 215–19.

______ **and** ______, "Team Production in Political Majorities," *Micropolitics*, 1982, *3*, 111–21.

Fenno, Richard, *Congressmen in Committees*, Boston: Little, Brown, 1973.

Ferejohn, John, *Pork Barrel Politics: Rivers and Harbors Legislation 1947–1968*, Stanford: Stanford University Press, 1974.

Fiorina, Morris P., *Representatives, Roll Calls, and Constituencies*, Lexington: D.C. Heath, 1974.

Kalt, Joseph and Zupan, Mark, "Capture and Ideology in the Economic Theory of Politics," *American Economic Review*, June 1984, *74*, 279–300.

Kau, James and Rubin, Paul, "Self-Interest, Ideology and Logrolling in Congressional Voting," *Journal of Law and Economics*, October 1979, *22*, 365–84.

Leavens, Donald R., "Legislator Compensation and The Determinants of Committee Action in the U.S. House of Representatives," unpublished doctoral dissertation, George Mason University, 1984.

Leibowitz, Arleen A. and Tollison, Robert D., "A Theory of Legislative Organization: Making the Most of Your Majority," *Quarterly Journal of Economics*, March 1980, *94*, 261–77.

Peltzman, Sam, "An Economic Interpretation of Congressional Voting in the Twentieth Century," *American Economic Review*, September 1985, *75*, 656–75.

Legi-Slate, On-Line Database, Washington, 1985.

U.S. Congressional Research Service, *Digest of Public General Bills and Resolutions*, part 1 and 2, 96th Congress 1st and 2nd Sessions, Washington: USGPO, 1981.

Product Line Rivalry: Note

By BIRGER WERNERFELT*

In an important paper, James Brander and Jonathon Eaton (1984) derive several results about equilibrium product configurations when multiproduct firms compete. However, their results are derived from the assumption that each firm in fact offers multiple (two) products. The purpose of this note is to shed some light on the conditions under which such product line competition will emerge. In addition, I consider whether an industry will supply standardized or differentiated products. For expositional clarity, I will conduct the analysis in the context of an example.

I. The Example

Brander and Eaton only allow their firms to produce four product types, which are specified as two pairs of substitutes. In order to investigate whether firms will offer product lines, I need to make it possible to offer a "standardized" product, while still preserving the incentives to sell more differentiated products. To do this, I adopt a Hotelling-style framework and allow firms to locate any number of products, at a cost d per product, anywhere in the interval $[0,1]$.

I then must specify a demand structure that incorporates the tension between standardization and differentiation. For this purpose, I assume the existence of two groups of consumers, indexed by 0 and 1, according to their preferred values of the product attribute α. Consumers are only willing to buy products whose α values differ at most by one-half from their preferred values, and they will demand a price discount on products with α values other than 0 or 1. In particular, I assume that the utility functions are such that if consumers in group 0 have a choice between several products ($j=2,3,\ldots$) located at $\alpha_j(\leq 1/2)$ and priced at p_j, they will select that which minimizes $p_j+\beta\alpha_j^2$ ($\beta>0$). Similarly, group 1 will minimize $p_j+\beta(1-\alpha_j)^2$, $(1-\alpha_j\leq 1/2)$. If inverse demand is linear in these "attribute-adjusted" prices, it follows that product j will sell in group 0 at

$$p_{0j}=10-bX_0-\beta\alpha_j^2,\quad \alpha_j\geq 1/2,\quad b>0,$$

where X_0 is the total volume purchased by group 0 and we have normalized the intercept at 10. If group 1 is identical (except for its preferences about α), and that group is sold X_1, the price will be $p_{1j}=10-bX_1-\beta(1-\alpha_j)^2$, $1-\alpha_j\leq 1/2$.

I model competition in this market as a two-stage game. In the first stage, each of n firms locates one or more products on $[0,1]$; in the second stage, these firms engage in Nash quantity competition with zero marginal costs. Given this dynamic structure, firms will make first-stage decisions with an eye to their implications in the second stage. To find the perfect equilibria of the game, therefore, I first solve the quantity game for all different location decisions and then evaluate them, given equilibrium play in the second stage. Since firms can locate any number of products over the entire interval $[0,1]$, the backwards solution procedure at first seems very complex. However, we can simplify matters a great deal. To see how, consider a product which is not located at 0, $1/2$, or 1. Such a product will only sell in one group and will command a price strictly below that of the appropriately targeted product (0 or 1); therefore, the only possible equilibrium locations are 0, $1/2$, and 1. Furthermore, it never pays for a firm to locate a product at $1/2$ and another one at 0 or 1 at the same time, since strictly higher revenues can be obtained from locating at 0 and 1 only. Roughly following Brander and Eaton's terminology, I refer to the case where each firm specializes in one end of the market as

*J. L. Kellogg Graduate School of Management, Northwestern University, Evanston, IL 60201.

TABLE 1—b TIMES PROFIT

	1/2	0	0+1
1/2	– $2(10-\beta/4)^2/9-bd$	$(10+\beta/4)^2/9-bd$ $(10-\beta/4)^2/4+(10-\beta/2)^2/9-bd$	$2(10+\beta/4)^2/9-2bd$ $2(10-\beta/2)^2/9-bd$
1	– –	– $25-bd$	$325/9-2bd$ $100/9-bd$
0+1	– –	– –	– $200/9-2bd$

segmentation, while interlacing describes situations where all firms offer two products. Standardization occurs when all products are located at 1/2.

For the special case of a duopoly, it is tedious but trivial to compute the payoffs for the first stage of the game, given equilibrium play in the second stage. For example, if one firm locates at 0 while the other firm locates at 1, they will be local monopolists and can each reap profits of $25/b-d$. Similarly, if both firms locate at 1/2, we have a homogeneous duopoly and in Nash equilibrium each firm gets profits of $2(10-\beta/4)^2(9b)^{-1}-d$. The entire payoff matrix, which obviously is symmetric, is shown in Table 1. From the table, it is trivial but tedious to establish that

I. (1/2,1/2)—standardization—is a perfect equilibrium if

$$\beta<\min\{9bd/20, 40(3-2\sqrt{2}, 40-12(bd/2)^{1/2}\}$$

II. (0,1)—segmentation— is a perfect equilibrium if

$$\beta>8(17-\sqrt{189})/5;\ 100/9<bd<25.$$

III. (0+1,0+1)—interlacing—is a perfect equilibrium if

$$\beta>20-(400-18bd)^{1/2};\ bd<100/9.$$

IV. (1/2,0+1) is a perfect equilibrium if

$$\beta<\min\{20-(18bd)^{1/2}, 20-(400-18bd)^{1/2}\};\ bd<20\beta/9.$$

V. There are no other perfect equilibria.

To interpret these conditions, it also may be helpful to refer to Figure 1. As one would expect, the basic forces at work are the incentives to differentiate, measured by the heterogeneity of tastes (β), and the costs of offering additional products, measured by the scaled fixed costs (bd). Assume first that

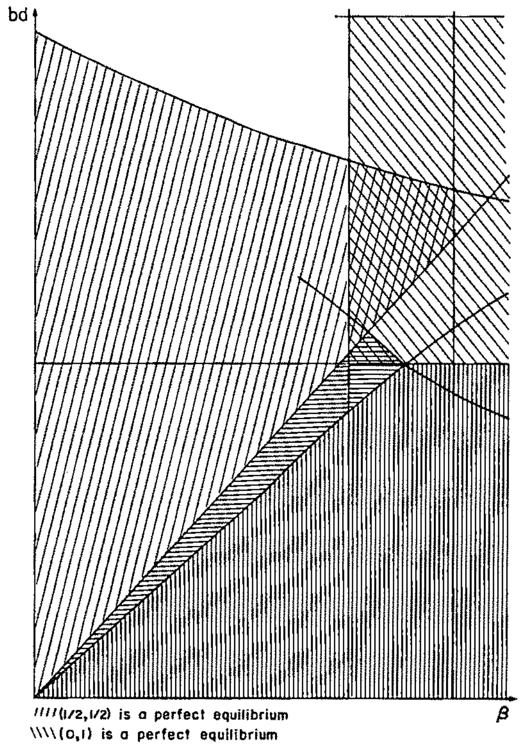


FIGURE 1

there are high fixed costs (so bd is high). If tastes are homogeneous (β is small), both firms will offer a single standardized product. If tastes are heterogeneous, the firms will segment the market and each cater to one group. In certain intermediate cases, the incentives to differentiate and the costs of differentiating balance out so that only one firm can do it profitably.

In the regions where there are multiple perfect equilibria, we can get uniqueness by letting the firms enter sequentially so that the first mover can impose the most profitable equilibrium. There are two such regions, one in which $(1/2,1/2)$ and $(0,1)$ are perfect equilibria, and one in which $(1/2,0+1)$ and $(0,1)$ are perfect equilibria. In both cases, it turns out that sequential play will lead to the configuration $(0,1)$ because this eliminates competition. (In this example, it is not advantageous to preempt the center of the market, since all demand is at the extremes.)

The assumption that only two firms compete is clearly unsatisfactory. One would expect continued entry until profits are driven to zero. If we concentrate on symmetric equilibria in industries with n firms, we can find the profits from standardization, segmentation, and interlacing as

$$b\Pi(1/2|1/2) = 2(10-\beta/4)^2/(n+1)^2 - bd$$

$$b\Pi(0|1) = 100/(1+n/2)^2 - bd$$

$$b\Pi(0+1|0+1) = 200/(n+1)^2 - 2bd.$$

From this it can be seen that the segmented structure allows the greatest number of firms, while standardization comes in second. As one would expect the interlaced product configuration is the best entry deterrent, a result also found by Brander and Eaton.

II. Concluding Remarks

In an effort to supplement the product line rivalry results obtained by Brander and Eaton, I have examined a model in which firms may or may not choose to offer multiple products. The objective is to identify those conditions under which multiple product (or "product line") offerings will emerge. The principal result is that heterogeneity in tastes and low product-specific fixed costs favor "product line rivalry" in the sense that both firms will produce two products and compete directly with each other. If tastes are more homogeneous and product-specific fixed costs are high, both firms produce a single standardized product. If both heterogeneity and fixed costs are high, each firm will produce a different single product.

Since this analysis has been conducted in the context of a specific example, the results have to be interpreted with caution. Generalization to a more realistic setting would complicate the exposition considerably. For example, if consumers are continuously distributed on $[0,1]$, it is no longer possible to establish that locations other than 0, 1/2, and 1 are dominated; therefore, the first stage of the game has to be analyzed as an infinite game. Nevertheless, I trust that the reader will find the intuition behind these results quite robust.

REFERENCE

Brander, J. A. and Eaton, J., "Product Line Rivalry," *American Economic Review*, September 1984, *74*, 323–34.

An Unpublished Letter from Malthus to Jane Marcet, January 22, 1833

By BETTE A. POLKINGHORN*

E I Coll Jan 22, 1833

My dear Madam

I have read John Hopkins's Notions on Political Economy with great interest and satisfaction, and am decidedly of opinion that they are calculated to be very useful. They are in many respects better suited to the labouring classes than Miss Martineau's Tales which are justly so much admired. I am strongly therefore inclined to advise you to publish them in as cheap a form as you can, for general circulation, and to give away. We shall be happy to purchase a dozen of them to distribute to the Cottagers in our neighborhood. I think your doctrines very sound, and what is a more essential point, you have explained them with great plainness and clearness. [If I were obliged to find any fault, I should say that you have presented in rather too brilliant and unshaded colours the advantages which would accrue from the abolition of the Corn Laws, so as to excite expectations which cannot be realized.[a] *In the actual state of the redundancy of labour in this country, it appears to me scarcely possible to conceive that the money wages of labour will not fall nearly in proportion to the price of corn, and the labourers be greatly disappointed.*[b] It will no doubt give a stimulus to foreign trade; but it must for a considerable time aggravate the redundancy of labour in country parishes; and during the process of the change, there will probably be more thrown out of work than in any other case of the restrictions of the freedom of trade, on account of the largeness of the concerns. It will also tend to raise the value of money and increase the pressure of the national debt. *Still I am for the removal of the restrictions*, though not without fear of the consequences.][c] I have been (interrupted?) and must finish.

Most truly yours
T. Robt Malthus

While doing research for a book on the life and work of Jane Marcet, I was fortunate enough to discover the above hitherto unknown letter written by Malthus in 1833, shortly before his death.[1] The recipient, Jane Marcet, was one of the leading popularizers of political economy in the first half of the nineteenth century. She wrote *Conversations on Political Economy* (1816) for adults and young people, *John Hopkins's Notions on Political Economy* (1833) for workers, and *Rich and Poor* (1851) for children. Her work was held in high esteem by her contemporaries such as J. B. Say, who called her "the only woman who had written on political economy and shown herself superior even to men" (*DNB*, 1899, p. 123). Writing from East India College, Hertfordshire, Malthus

*Department of Economics, California State University, Sacramento, 6000 J Street, Sacramento, CA 95819.
I thank Brinley Thomas and two anonymous referees for helpful comments on an earlier draft of this paper.

[1] The letter is in the Marcet Collection of the Archive Guy de Pourtalès. It is published with the permission of the Fondation Guy de Pourtalès-Etoy, et Centre de Recherches sur les Lettres Romands, Université de Lausanne, Switzerland. It is postmarked January 23, 1833 and addressed: Mrs. Marcet, 49 Weymouth Street, Portland Place, London. The underlining (shown here in italics) and brackets are by Malthus.

Jane Marcet was of British birth and Swiss descent. In London in the fall of 1983, I attempted to locate descendants of Mrs. Marcet so as to determine if papers still existed. The search at one point involved the perusal of each of the 283 telephone books in Britain in an effort to identify families with the surname Pasteur—the family name of some of Jane Marcet's descendants. I wrote to every person found with that name and, through the process, was able to identify the people I sought. They had papers and were friendly and cooperative. They informed me that there were many more papers in Switzerland. Therefore, I went to Switzerland and was given permission to study these records. While reading in the Marcet Collection of the Fondation Guy de Pourtalès, I discovered a package of letters written about Jane Marcet's books. The letter from Malthus, unknown to modern scholars, was included in this package.

was acknowledging his receipt of Mrs. Marcet's *John Hopkins's Notions on Political Economy*. The letter deals with the Corn Law controversy and ends with an important sentence—"*Still I am for the removal of the restrictions, though not without fear of the consequences.*"

I. Footnotes to the Letter

[a]Marcet's opinion was typical of opponents of the Corn Laws—food would be cheaper and workers' real income higher with free importation of corn.

[b]Although it sounds strange to a modern economist, Malthus thought the working class better off when the price of corn was high rather than low. Malthus is correct if it is assumed that money wages vary with the price of corn and that the prices of other goods do not vary as much. For a thorough explanation, see William Grampp (1956).

[c]In his *Observations on the Effects of the Corn Laws* (1814), Malthus had set out objectively the case for and against the Corn Laws without committing himself. As he put it in his subsequent 1815 pamphlet, "Some of my friends were of different opinions as to the side, towards which my arguments most inclined. This I consider as a tolerably fair proof of impartiality" (p. 138). In this pamphlet, Malthus took into account "some facts which have occur[r]ed during the last year, and which have given, as I think, a decisive weight to the side of restrictions" (p. 138). He referred to the sharp fall in the price of corn and in the price of bullion, and the decision of France to free the export of corn. His conclusion was as follows: "I firmly believe that, in the actual state of Europe, and under the actual circumstances of our present situation, it is our wisest policy to grow our own average supply of corn..." (p. 173). The weights which Malthus attached to the pros and cons were clearly influenced by his evaluation of contemporary events.

The case for restrictions was elaborated in *Additions To The Fourth and Former Editions of An Essay On The Principle of Population* (1817) and in *Principles of Political Economy Considered with a View to Their Practical Application* (1820). While arguing that free trade was best for the world as a whole, Malthus claimed that England's long-run interests required a balanced growth of the agricultural and manufacturing sectors, with neither preponderating over the other: "Whether a balance between the agricultural and commercial classes of society, which would not take place naturally, ought, under certain circumstances, to be maintained artificially, must appear to be the most important practical question in the whole compass of political economy" (1817, p. 182).

The key sentence in Malthus' letter to Marcet, "*Still I am for the removal of the restrictions*, though not without fear of the consequences," indicates that in the circumstances of 1833 he had changed his weighting of the various arguments. I can only speculate on the reasons. There may be a clue in the fact that in the previous year, in a letter to Thomas Chalmers (1832), he wrote, "I quite agree with you in regard to the moral advantage of repealing the Corn Laws." Chalmers' argument was that repeal would lessen "the burning discontent all over the land" and "sweeten the breath of British society," and that it was immoral for a government dominated by landlords to use the power of the state to maintain dear food. Moreover, the advantage of cheap food for the working class would not be lost if population growth was checked by moral restraint (1832, pp. 523–24 and 539–40). It is possible that these considerations tipped the balance for Malthus in 1833. To the best of my knowledge, this was his last word on the subject. He died on December 29, 1834.

REFERENCES

Chalmers, Thomas, *On Political Economy In Connection With the Moral State & Moral Prospects of Society*, Glasgow: William Collins, 1832; Reprinted, New York: Augustus Kelley, 1968.

Grampp, William, "Malthus on Money Wages and Welfare," *American Economic Review*, December 1956, *46*, 924–35.

Malthus, Thomas Robert, *Observations on the*

Effects of the Corn Laws, 2nd ed., London: J. Johnson and Co., 1814; reprinted in *The Pamphlets of Thomas Robert Malthus*, New York: Augustus Kelley, 1970.

______, *The Grounds of An Opinion on the Policy of Restricting the Importation of Foreign Corn*, London: John Murray, 1815; reprinted in *The Pamphlets of Thomas Robert Malthus*, New York: Augustus Kelley, 1970.

______, *Additions To The Fourth and Former Editions Of An Essay On The Principle of Population*, London: John Murray, 1817.

______, *Principles of Political Economy Considered with a View to Their Practical Application*, London: John Murray, 1820.

______, Letter of March 6, 1832, to Thomas Chalmers, Thomas Chalmers Collection, New College Library, Edinburgh, Scotland.

[Marcet, Jane], *Conversations on Political Economy in which the Elements of That Science are Familiarly Explained*, Longman, Hurst, Rees, Orme and Brown, London: 1816.

______, *John Hopkins's Notions on Political Economy*, London: Longman, Rees, Orme, Brown, Green and Longman, 1833.

Marcet, Mrs., *Rich and Poor*, London: Longman, Brown, Green and Longman, 1851.

Dictionary of National Biography **(*DNB*)**, London: 1899, *36*, 122–23.

Competitive Payments Systems: Comment

By ROBERT L. GREENFIELD AND LELAND B. YEAGER*

In the September 1984 issue of this *Review*, Lawrence H. White criticizes several recent proposals for monetary reform, including what we (1983) call the BFH system. Under this system, the government, precluded from issuing money, would merely define a new unit of account and encourage the unit's general adoption by using it in all its own, pricing, contracting, and accounting. The definition would run in terms of so comprehensive a bundle of precisely gradable items with continuously quoted prices that the new unit would have a stable general purchasing power.

According to White, our considering that the system's unit of account has operational meaning merely means that we "find the concept of keeping track of relative prices by use of a numeraire unit not incoherent or self-contradictory" (p. 703). But we see in the unit's operationality more than just the calculations of Walras' auctioneer. We use the term "operational" only after having satisfied ourselves that in the course of honest-to-goodness market activity, the unit denominating privately issued notes and deposits would pose no threat of prying itself loose from its commodity-bundle definition. The absence of any dominant medium of exchange, or "cashlessness," to use White's term, ensures their adhesion.

The danger of the commodity-bundle-defined unit of account falling into desuetude would exist if all things serving as media of exchange were essentially indistinguishable from one another. Quantities of a homogeneous medium of exchange would be measurable in a common unit distinct from the actual commodity bundle; and that common unit, the medium-of-exchange unit, might indeed rival the commodity bundle itself as the unit of account. But under the BFH system, which would eschew government money, no such single homogeneous medium of exchange would exist, so no tension could arise between regarding a unit of it and regarding the commodity-bundle-defined unit as the unit of account. Notes and demand deposits, as well as checkable *equity* holdings in the institutions issuing them, would be distinguishable by their private issuers, who individually would face competitive pressures to keep their obligations *meaningful*.

These pressures would come to a focus at the clearinghouse. Convenience would dictate, however, that in settling clearinghouse balances due on account of notes and checks (checks drawn on both deposits and equity holdings), issuers transfer not quantities of the standard bundle itself, but redemption property *worth* as many standard bundles as the number of units to be settled. Issuers would thus keep their obligations meaningful by making them "indirectly" convertible, convertible into "bundles-worths," and so prevent the unit of account from losing contact with its commodity-bundle definition.

While retaining and enhancing the advantages of a single definite pricing and accounting unit and convenient methods of payment, the BFH system would avoid the absurdity of a unit of account whose size is the supply-and-demand-determined value of any medium of exchange. With the unit of account and media of exchange separated, the unit's value would be established by definition, leaving the quantities of the various media of exchange directly responsive to the demands for them. No longer could there be too much money, causing inflation, too little money, causing depression, or a temporal sequence of imbalances, causing stagflation.

Somehow thinking that we expect unit-of-account-denominated deposits to disappear even while outside money continues to

*Fairleigh Dickinson University, Madison, NJ 07940 and Auburn University, Auburn, AL 36849, respectively.

exist, White emphasizes today's "comparative costliness of check writing against money market funds" (p. 707). The BFH system, however, would dispense with outside money, from whose dominance that cost differential—if indeed one actually exists—might spring. People could choose between checkable equity holdings and checkable unit-denominated deposits. We conjecture that the former would play a prominent role, but the logic of the BFH system in no way presupposes their displacing unit-denominated deposits.

White (p. 710) says that deregulation does not imply the disappearance of outside money. Needless to say, it does not. But not only does he seem to say (p. 712) that we think it does, he also considers "cashlessness" itself to do some violence to the very institution of credit. In White's view (p. 710), the BFH system is circular somehow because bonds and interest on them are supposedly payable, ultimately, only in shares in portfolios of bonds. (Actually, we expect checks on equity funds to be used in payments also, but that is a side point only.)

What trouble does the supposed circularity cause? If White thinks the BFH system suffers from circularity, what does he think of our existing system, in which bonds are payable only in money—pieces of paper not payable in anything at all? Is White consoled because those pieces of paper count as outside money? In both the BFH system and our existing system, bondholders receive interest and repayment of principal in property, typically paper assets, that they can exchange for desired goods and services. (The processes that give actual exchange value to this payment property differ between the two systems, but that is a minor detail insofar as White's tacitly taking redeemability as the criterion of noncircularity is concerned.) The longer people wait to consume goods and services, the more they can ultimately consume. That is what interest is about. Where is the difficulty?

White (p. 710) explains how money originated and how the unit of account has been linked to the medium of exchange. He says that the BFH system could not evolve in the same spontaneous fashion, as if this fact counted heavily against it. But why should an account of historical evolution, or conjectures about what might have spontaneously evolved or might still evolve, form the centerpiece of judgments about what is desirable now? Accumulated experience, new ideas for alternative systems, and advanced technology of communications, calculations, and record-keeping open up new possibilities. Why not take them into account in deciding where to proceed from here?

Any monetary reform must begin from where we are now. Dismantling government domination of the existing system will require deliberate policy actions, and the particular actions taken will unavoidably condition the successor system. Its capacity for emerging spontaneously is a spurious criterion of desirability.

REFERENCES

Greenfield, Robert L. and Yeager, Leland B., "A Laissez-Faire Approach to Monetary Stability," *Journal of Money, Credit and Banking*, August 1983, *15*, 302–15.

White, Lawrence H., "Competitive Payments Systems and the Unit of Account," *American Economic Review*, September 1984, *74*, 699–712.

Competitive Payments Systems: Reply

By LAWRENCE H. WHITE*

My 1984 article critically reviewed the work of several authors who have conceived of competitive payments systems devoid of outside money and free of any link between payment media and the unit of account. I concluded that cashlessness, and the divorce of the unit of account from own-units of payment media, are not natural products of unrestricted competition. Further, some imagined systems incorporating these features suffer from internal incoherence. Robert Greenfield and Leland Yeager (1986) register three principal complaints about my treatment of their 1983 contribution to this literature. First, I did not appreciate just how operational they think the unit of account in their system (what they call a "BFH system") really is. Second, I made an argument, concerning the circularity of bonds payable only in bonds, which they find difficult to understand. Third, and evidently most importantly, they believe that I argued "as if" to question the desirability of their system as a serious proposal for monetary reform. I will respond to each of these complaints in turn.

I. Operationality

The issue of whether the unit of account in a Greenfield-Yeager system is operational, that is, fit for proper functioning, can be framed in various ways. The narrowest of these is whether a Walrasian auctioneer could coherently use the unit. Greenfield and Yeager apparently agree with me that an affirmative answer to that question does not establish very much, for they insist (justifiably) that they have framed the issue more broadly than that. They argue that "honest-to-goodness" decentralized trading would not pry the unit denominating payment accounts (demand deposits or checkable equity holdings) away from a commodity-bundle definition of the unit of account initially adopted. No divorce could come about, they say, because the only plausible alternative unit for denominating payment accounts would be some quantity of a common medium of exchange, and their system has no common medium of exchange.

Let us suppose, for the sake of argument, that a Greenfield-Yeager system has been established. Would market forces subsequently promote the emergence of a common medium of exchange, or would they prevent such a development? As far as I can tell, Greenfield and Yeager have not addressed this question. (Nor have I previously.) They do state that "under [our] system, which would eschew government money, no such single homogeneous medium of exchange would exist" (p. 848). But surely absence of government money does not insure absence of *any* homogeneous outside money: a specie standard with private mints and no government bank furnishes a conceptual counterexample.[1]

In Carl Menger's theory of the origin of money, which my article recounted, the needs of hand-to-hand traders promote the emergence of a common medium of exchange which eventually takes the form of an outside currency. For present purposes, however, I grant the assumption that in a sophisticated payments system the public can happily do without an outside currency. Market pressure for a common medium of exchange would instead be felt most strongly at the clearinghouse. (For simplicity, assume

*New York University, New York, NY 10003. I am thankful to members of the Austrian Economics colloquium at NYU for comments and discussion.

[1] Greenfield and Yeager (1986) add that payment account holdings would be nonhomogeneous, differing according to issuer. But this is logically unrelated, as the same counterexample shows, to the existence or nonexistence of a homogeneous outside money.

that a single clearinghouse covers the entire economy.) Greenfield and Yeager speak of clearing balances being settled by transfer of nonhomogeneous "redemption property" (1986, p. 848; 1983, p. 307). They do not further elaborate except to speculate that the settlement assets agreed upon by member funds might consist of "specified securities." Already this recognizes the crucial point that not all assets are equally acceptable to all traders as payment.

How then are settlement assets and their values agreed upon? Certainly it would not be feasible to negotiate each settlement individually. Suppose instead that there is a pre-approved list of specified securities. Who chooses the securities to be remitted in a particular day's settlement? It does not seem workable to let the paying member fund make the choice. Assuming end-of-day settlement using securities evaluated at closing prices, the fund would have an incentive to remit securities which up-to-the-minute news indicated would be likely to lose the most value between that day's (or the most recent market day's) closing and the next market day's opening. And it could profit the fund to bid up a closing security price in order to lock in an artificially high price at which to unload a great quantity of the security (including the quantity it had just purchased). There would be even graver problems with evaluation and choice of securities for purposes of intraday settlement, when the spread between bid and ask prices is obvious. For these reasons, there would be market pressure for a homogeneous settlement asset. The clearinghouse can provide such an asset by holding member fund redemption property on account, and pooling it, giving each fund homogeneous shares in the clearinghouse portfolio (hereafter *CP*). The clearinghouse can then make settlement instantaneously by transfer of *CP* shares between accounts. That ability is important in light of the fact that wire transfers account for approximately three-fourths of all transactions in the United States today in unit-of-account volume (Maxwell Fry and Raburn Williams, 1984, p. 6).

This arrangement makes the participating funds themselves owners of shares in a funds' fund, just as clearing banks today own deposits at a bankers' bank. More importantly, it gives *CP* shares many of the characteristics of outside money. The *CP* shares are effectively a redemption medium for transfers among ordinary commercial payment accounts, just as deposits at the Federal Reserve are in the present American banking system. Over-the-counter redeemability might also emerge. *CP* shares are routinely accepted as a medium of exchange, because no one will refuse to accept the ultimate clearing asset. A question requiring further thought is whether *CP* shares can be spent into existence, and, if so, with what consequences.

Does the *CP* share constitute a unit which could rival the government-chosen commodity bundle as a unit of account? It does if the clearinghouse defines the *CP* share in "physical" terms, for example, one *CP* share equals 1.0 shares Alcoa common stock, 2.5 shares Burlington Northern stock, 1.7 shares Conoco stock, and so on down a list. The price of a *CP* share in terms of the government-stipulated numeraire would then vary day to day. Correspondingly, the *CP* share price of the commodity bundle to which all government accounting and obligations were indexed would vary from day to day. If, on the other hand, the clearinghouse were to denominate *CP* shares in unit-of-account terms, just as the typical money market mutual fund today fixes its share price at one dollar, then no rivalry would exist. The clearinghouse might well choose not to do so, however, in order to avoid the awkwardness of posting a price for acquisition and surrender of *CP* shares by funds, which could only be paid in bundles of primary securities, in units other than the units it would routinely accept and pay.

II. Circularity

I argued that there is a circularity problem in a Greenfield-Yeager system if two conditions simultaneously hold: 1) bonds are exclusively claims to streams of payment in fund shares; and 2) fund shares are claims to portfolios consisting exclusively of bonds. The problem may perhaps be grasped more

clearly by considering the absurdity of consols which are exclusively claims to future streams of similar consols.[2] Who would want to buy such a claim? A transactor attempting to value it faces an infinite regress. I don't know how to make the difficulty any clearer. My conclusion was not that *every* cashless competitive payments system is inherently circular, but that to avoid circularity, conditions 1 and 2 could not both hold. Circularity would be avoided if either bonds were claims to streams of commodities or equities, or fund portfolios consisted of equities. (I argued that either of these arrangements would introduce other problems, however.) Present-day bonds payable in fiat money pose no circularity problem. To recognize this is not to find any great solace in fiat money's irredeemability. The apparent "bootstrap" paradox of fiat money itself having a positive value, I argued (p. 706), can be resolved by understanding the historical transition from redeemable to irredeemable central bank liabilities.

III. Reform

Contrary to the first sentence of Greenfield and Yeager's comment, I did not treat their 1983 piece as a proposal for monetary reform. I therefore did *not* intend my conjectures about spontaneous evolution to "form the centerpiece of judgments about what is desirable now," as they suppose (p. 000). Instead, I deliberately and explicitly limited my critique of their system, and of the cashless competitive payments systems of Black and Fama, to questioning "their applicability for modeling current arrangements or predicting future arrangements" (p. 706). I thought that a purely analytical approach was consistent with the approach of Greenfield and Yeager, who in their opening paragraph remark: "Regardless of who if anyone may actually advocate the system, contemplating it is instructive. It illuminates, by contrast, some characteristics of our existing and recent systems" (1983, p. 302). I agree fully with that. And I would add that most readers will probably find the idea of a cashless competitive payments system easier to contemplate seriously when it is presented as an analytical construct than when it is presented as a reform proposal. But I am willing to deviate from my original purely analytical orientation in order to address briefly here the issues raised by cashlessness as a reform proposal.

I do not at all wish to question Greenfield and Yeager's preference for "Dismantling government domination of the existing system" (p. 849), that is, for moving to a private and unregulated payments system. (For what it is worth, I share that preference.) We agree that doing this entails deregulating banks and other financial institutions. There remains, however, the question of how to undo government's current control over the quantity of basic money. As they correctly insist, any approach requires deliberate policy actions that will condition the successor system. One avenue of reform (the one I happen to favor) is to take steps to enable private competitively issued money to supplant government fiat money. Commodity money, having historical precedent, is the most obvious form private outside money might take, but noncommodity monies as imagined by Hayek (1978) are also worth consideration. Greenfield and Yeager's alternative avenue is simply to abolish money.

In its starkest outlines, Greenfield and Yeager's argument for reform runs as follows. 1) Monetary payments systems inherently have important features which are socially undesirable. 2) Therefore it is desirable to abolish money. In advancing these two propositions their argument reminds me of S. Herbert Frankel's characterization of Keynes' outlook: "[I]t rests on the fear of money itself.... Keynes...sees money as distorting everything and wants the authority of the state to force money to reflect a less disturbing image" (1977, p. 3). Greenfield and Yeager, while fearing money, instead want the state to facilitate the abandonment of money. They propose that an effective and desirable way to abolish money is to have government announce and use a unit of account defined in terms of a bundle of

[2] I borrow this example from Kevin D. Hoover (1985, p. 55).

goods so wide as to be totally unusable as a medium of exchange.[3]

The logical gap between steps 1 and 2 should be obvious. Granting that the use of money carries with it certain social costs (foregone benefits of barter) does not compel one to conclude that its costs outweigh its benefits. One of my purposes in tracing the spontaneous evolution of money, and in emphasizing the supreme saleability of money, was to indicate that there are important benefits to using a common medium of exchange, namely in facilitating transaction. These benefits are never mentioned by Greenfield and Yeager, and seem to have been overlooked.[4] Such an oversight is surprising given that Yeager is the author of a classic account of "the essential properties of a medium of exchange" which emphasizes money's supreme saleability in comparison with other assets. In that paper, Yeager recognizes that money has uniquely low transactions costs, and explains that for an asset to have "the lowest transactions costs" means that "loosely speaking, it is the most convenient medium of exchange" (1968, p. 67). Surely the extra convenience of using money—of having a generally accepted asset or ultimate settlement—is a genuine benefit that ought not to be neglected in the evaluation of monetary vs. nonmonetary payments systems.

Most of the advantages that Greenfield and Yeager (1983, pp. 308–11) claim for their system may be attained, I believe, without abolishing money. Reasonable stability in the purchasing power of the numeraire, an end to inflationary finance, competitive innovation in payments institutions, resistance to financial panics, and mitigation of macroeconomic difficulties through a demand-elastic supply of particular forms of payment media, would all be promoted by deregulation of banking (including the private issue of currency) combined with freezing or denationalizing outright the supply of base money.[5] In addition, either freezing or denationalizing the monetary base (the latter by redeeming fiat dollars for some commodity presently stockpiled by government) avoids an important disadvantage of cashlessness: the transition to cashlessness implies significant wealth losses to relatively heavy base-money holders.

[3]Greenfield and Yeager (1983, p. 303) explicitly eschew state force against money users. Presumably, however, citizens are to be forced to pay taxes in commodity-bundle-denominated media. It is nonetheless far from obvious, to anyone skeptical of the state theory of money, that these measures would be sufficient to make private traders abandon dollars in order to adopt the new system.

[4]Greenfield and Yeager do assert that their system would retain "convenient methods of payment" (1986, p. 848), for example, check writing, but they evidently see no convenience in a common *medium* of payment.

[5]George Selgin (1986) provides detailed arguments for these results, particularly the first, fourth, and fifth.

REFERENCES

Frankel, S. Herbert, *Money: Two Philosophies*, Oxford: Basil Blackwell, 1977.

Fry, Maxwell J. and Williams, Raburn M., *American Money and Banking*, New York: Wiley & Sons, 1984.

Greenfield, Robert L. and Yeager, Leland B., "A Laissez Faire Approach to Monetary Stability," *Journal of Money, Credit and Banking*, August 1983, *15*, 302–15.

______ and ______, "Competitive Payments Systems: Comment," *American Economic Review*, September 1986, *76*, 848–49.

Hayek, F. A., *The Denationalisation of Money*, 2nd ed., London: Institute of Economic Affairs, 1978.

Hoover, Kevin D., "Causality and Invariance in the Money Supply Process," unpublished doctoral dissertation, Oxford University, 1985.

Selgin, George A., "The Theory of Free Banking," unpublished doctoral dissertation, New York University, 1986.

White, Lawrence H., "Competitive Payments Systems and the Unit of Account," *American Economics Review*, September 1984, *74*, 699–712.

Yeager, Leland B., "Essential Properties of a Medium of Exchange," *Kyklos*, 1968, *21*, 45–69.

The Evans and Heckman Subadditivity Test: Comment

By TOSHIYUKI SUEYOSHI AND PETER C. ANSELMO*

David Evans and James Heckman's article in this *Review* (1984) described the development of a numerical test to measure subadditivity of a cost function with an application to the Bell System. Unfortunately, a feature of this test was not discussed and its related figures for admissible regions are not drawn correctly in their article. (See also their 1983 papers.) In this comment we briefly mention the feature and correctly draw two figures for admissible regions for subadditivity measurement.

An important feature of the Evans-Heckman test is the restriction of the subadditivity test to years for which the outputs of the multiple products are at least twice the lowest levels of their products, so that their multiple imaginary firms may provide multiple products whose total quantities sum to the same output as would be found in the case of single-firm production.

Another unique feature of the Evans-Heckman test is the inclusion of two constraints related to the output ratio; R_L, the lower bound and R_U, the upper bound. The constraints are expressed as equations (10) and (11) in their article. They treat a restricted region as an admissible region in which all the entries are between R_U and R_L. Other entries, either above R_L or below R_U, are eliminated from subadditivity measurement. (See their Figure 2, p. 618.) However, as the symmetry property indicates, the "real admissible region" to measure subadditivity is required to satisfy both the restricted region as they proposed and the symmetry condition that we here provide.

Figure 1 illustrates how their proposed constraints restrict a measurable region. The smallest hypothetical production size is q_M. The curve from a to i denotes output configurations from the first to last observations. $\tilde{q}_t$ is an output vector for a certain year t whose subadditivity may be measured in the rectangular region (a-b-c-d-a). The broader measurable region is divided into a restricted

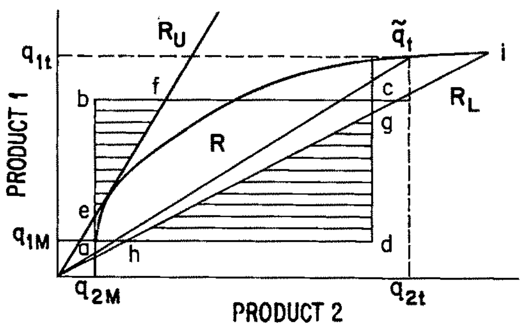


FIGURE 1. DETERMINATION OF RESTRICTED REGION (R)

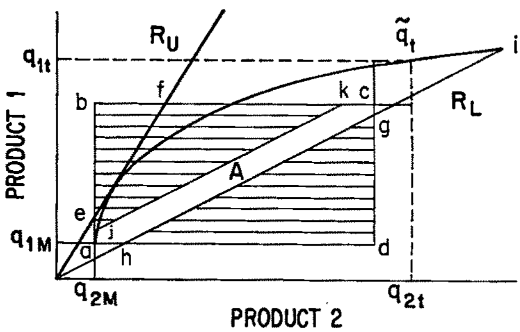


FIGURE 2. REAL ADMISSIBLE REGION (A)

*Assistant Professor, Ohio State University, School of Public Administration, 1775 College Road, Columbus, OH 43210, and Ph.D. candidate, University of Texas-Austin, Graduate School of Business, Management Department, Austin, TX 78712. This research was partly supported by the Management Department, University of Texas at Austin. We thank William W. Cooper for helpful comments and encouragement.

region (R: a-e-f-c-g-h-a) whose entries all satisfy the two constraints, R_L and R_U, and the other region in which the constraints are not satisfied. The latter region consists of two triangles (e-b-f-e and h-g-d-h) in Figure 1.

On the other hand, the admissible region for subadditivity measurement is illustrated in Figure 2. The admissible region is expressed by the area A: (a-j-k-c-g-h-a) in which the distance ck and cg are equal to ah ($= ck$) and aj ($= cg$), respectively. The two constraints and the symmetry property diminish the measurable region for subadditivity as shown in Figure 2.

REFERENCES

Evans, David S. and Heckman, James J., "A Test for Subadditivity of the Cost Function with an Application to the Bell System," *American Economic Review*, September 1984, *74*, 615–23.

______ **and** ______, "Natural Monopoly," in D. S. Evans, ed., *Breaking Up Bell*, New York, North-Holland, 1983, 127–56.

______ **and** ______, "Multiproduct Cost Function Estimates and Natural Monopoly Tests for the Bell System," in D. S. Evans, ed., *Breaking Up Bell*, New York, North-Holland, 1983, 253–82.

ERRATUM

A Test For Subadditivity of the Cost Function with an Application To the Bell System

By DAVID S. EVANS AND JAMES J. HECKMAN*

In our paper published in this *Review*, September 1984, we presented a new test of the subadditivity of industry cost functions and applied the test to cost data from the Bell System. The data and cost functions used to perform the test were presented in our earlier paper (1983a). Unfortunately, there were some typographical errors as well as minor factual errors in both articles which make it difficult to replicate our empirical results precisely.

Here we present the correctly labeled cost functions fit on the correct data, and revised subadditivity tests for the revised cost functions. These corrections have no substantive effect on our finding that the Bell System was not a natural monopoly.

Some of the data reported in Table 10.14 of our paper (1983a) are incorrect. Because of typographical errors some of the reported data are not the correct data we used in our estimation. In addition, some of the data we used in our estimation were not copied accurately from the Bell System data. The major error occurs for 1977. Minor errors occur in some variables for 1964, 1970, and 1973. A table available from the authors reports the corrected data.[1] As it turns out, the data errors do not affect the estimated cost function. The cost and cost share data are calculated using 1967 prices. However, we normalize the input prices used as regressors in the cost and cost share equations to 1961 dollars. The discrepancy in the base used for the cost and input prices thus only affects the estimated intercept of the cost function, because we adopt a logarithmic specification for the cost function.

A more serious error that inhibits replication of our results is an inadvertent switch in the names of certain variables. Three multiplicative constants were also misstated so that the reported coefficients do not directly accord with the coefficients of the cost function. The correctly labeled cost function estimated on the corrected data is presented in Table 1. (This corresponds to Table 3 in our 1984 article.) These estimates are secured by imposing symmetry, homogeneity, and the cross-equation restrictions of producer theory despite the fact that these restrictions are rejected by the data.

In our papers (1983a; 1984), we use our estimated cost function to test for subadditivity of the Bell System technology for a two-firm breakup of the company (assuming that both firms have access to the same technology). In order to avoid extrapolation error, we define an admissible region in such a way that (a) no hypothetical firm can produce less of either local or toll output than is observed in the data, and (b) so that both hypothetical firms produce output within the range of *ratios* of the outputs actually observed in the data. Our test treats the two hypothetical firms symmetrically since they are assumed to have access to a common technology. The admissible region, so defined, does not necessarily contain any

*Fordham University, Bronx, NY 10458 and University of Chicago, Chicago, IL 60637. We thank Orley Ashenfelter, Bo Honore, and George Yates for their assistance in preparing this erratum.

[1] There are a few errors. The corrections for the cost data are: 1964 = 10542.5; 1970 = 16516.9; 1973 = 21190.3; 1977 = 34745.3. Toll output for 1947 = .34642; for 1948 = .37201. Price of capital 1977 = 1.41935. Research and Development 1966 = 1.66877. Share of capital in cost 1977 = .46712. Share of labor in cost 1977 = .39259. Many of these errors are typographical errors in our tables which were not made in processing the data.

TABLE 1—PARAMETER ESTIMATES FOR TRANSLOG COST FUNCTION CORRECTED FOR SERIAL CORRELATION

Incorrect Label	Parameter (Correct)	Estimate	Standard Error
–	Constant	9.054	.005
–	Capital (Price)	.536	.008
–	Labor (Price)	.354	.007
–	Local (Output)	.206	.299
–	Toll (Output)	.504	.219
–	Technology[a]	−.201	.086
–	Capital^2	.223	.026
–	Labor^2	.174	.027
–	Capital·Labor	−.183	.020
Toll^2	Local^2[a]	−16.646	4.310
Local^2	Toll^2[a]	−8.969	2.595
–	Local·Toll	12.167	3.105
–	Technology^2	−.180	.517
Capital·Toll	Capital·Local	.343	.140
Capital·Local	Labor·Local	−.362	.123
Labor·Toll	Capital·Toll	−.180	.085
Labor·Local	Labor·Toll	.161	.072
–	Capital·Technology	.081	.054
–	Labor·Technology	−.052	.048
Toll·Technology	Local·Technology	−1.553	1.503
Local·Technology	Toll·Technology	1.430	1.431
–	Autocorrelation Parameter for:		
	Cost Equation	.186	.105
	Share Equation	.706	.093

Summary Statistics	R^2	Degrees of Freedom	$D\text{-}W$[b]
Cost Function	.9999	14	2.15
Capital Share	.9753	27	1.57
Labor Share	.9834	27	1.61
System $MSE = 2.893$			

[a]In our 1983 and 1984 papers, the coefficients on Local^2 and Toll^2 should be multiplied by 2 and the coefficient on technology multiplied by 1/2 to obtain coefficients consistent with the parameterization of the cost equation given in our previous papers.

[b]This is the Durbin-Watson for the residuals from the ρ transformed equations.

particular historical output level for the Bell System.

We define the degree of subadditivity by

$$\text{Sub}_t(\phi,\omega)=\frac{\tilde{C}_t-\tilde{C}_t^A(\tilde{q}_t^A)-\tilde{C}_t^B(\tilde{q}_t^B)}{\tilde{C}_t},$$

where $\tilde{C}_t$ denotes predicted cost in year t at output level q_t, $\tilde{q}_t^A$ and $\tilde{q}_t^B$ are hypothetical output levels for firms A and B, respectively, and $\tilde{C}_t=\tilde{C}(\tilde{q}_t^A+\tilde{q}_t^B)$, $\tilde{q}_t^A+\tilde{q}_t^B=q_t$ and $\tilde{q}_t^A$ and $\tilde{q}_t^B$ satisfy the constraints on admissible outputs. If this expression is negative, the cost function is subadditive at $\tilde{q}_t^A, \tilde{q}_t^B$. If the expression is positive, the cost function is superadditive. Statistically significantly positive values reject the hypothesis of subadditivity or natural monopoly. Our test is local. Failure to find subadditivity within the admissible region is informative in rejecting the hypothesis of subadditivity.

We use our cost and cost share data to compute the revised estimates of $\text{Sub}_t(\phi,\omega)$ reported in revised Table 2. There are only small differences between the numbers in

TABLE 2—MAXIMUM PERCENT GAIN FROM MULTIFIRM VS. SINGLE-FIRM PRODUCTION[a]

Year	Percent Gain	Standard Error
1958	12.0	15.2
1959	17.3	14.0
1960	19.2	13.6
1961	21.7	13.7
1962	26.6	12.9
1963	32.0	12.9
1964	39.7	13.3
1965	46.4	15.7
1966	51.4	18.0
1967	55.0	20.3
1968	56.0	21.8
1969	54.4	24.7
1970	54.7	25.7
1971	52.4	26.1
1972	46.9	27.5
1973	43.3	20.0
1974	44.2	19.9
1975	47.0	19.7
1976	49.9	19.3
1977	52.4	18.8

[a]Entries equal max $\mathrm{Sub}_t \times 100$. A positive number indicates that multifirm production is more efficient than single firm production. The cost function is as reported in Table 2. Complete copies of subadditivity tables for each year are available on request from the authors.

Table 2 and the corresponding numbers in our original Table 2 (1984).[2]

[2]The labeling and scale multiplication errors noted in our discussion of Table 1 in our 1984 paper did not affect the estimates of $\mathrm{Sub}_t(\phi, \omega) \times 100$ reported, nor did the typographical errors noted in fn. 1 in this paper. SAS is used to perform the estimation. Some of the discrepancy between the estimates reported in this paper and the estimates reported in our previous papers is due to revisions in the SAS computer code.

REFERENCES

Evans, D. and Heckman J., (1983a) "Multiproduct Cost Function Estimates and Natural Monopoly Tests For the Bell System," in D. Evans, ed., *Breaking Up Bell*, Amsterdam: North Holland, 1983.

______ **and** ______, (1983b) "Natural Monopoly," in D. Evans, ed., *Breaking Up Bell*, Amsterdam: North Holland, 1983.

______ **and** ______, "A Test For Subadditivity of The Cost Function With an Application to the Bell System," *American Economic Review*, September 1984, *74*, 615–23.

ERRATUM

New Estimates of the Value of Federal Mineral Rights and Land

By Michael J. Boskin, Marc S. Robinson, Terrance O'Reilly, and Praveen Kumar*

In our paper, published in this *Review*, December 1985, Table 8 (p. 934) was printed with an incorrect heading stating the numbers in the tabulation were for "(Millions of Acres)": The correct heading is *Thousand of Acres*. This does not affect any other numbers in the paper; the calculations were done properly.

*Boskin: Professor of Economics, Stanford University, Stanford, CA 94305 and Research Associate, National Bureau of Economic Research; Robinson: Assistant Professor of Economics, University of California-Los Angeles, Los Angeles, CA 90024; O'Reilly and Kumar: graduate students, Economics Department, Stanford University.

Preliminary Announcement of the Program

ANNUAL MEETING
THE AMERICAN ECONOMIC ASSOCIATION

New Orleans, Louisiana, December 27–30, 1986

Saturday, December 27, 1986

10:00 A.M. EXECUTIVE COMMITTEE MEETING

Sunday, December 28, 1986

8:00 A.M. U.S. PRODUCTIVITY GROWTH: RESEARCH RESULTS FROM THE BUREAU OF LABOR STATISTICS
Presiding: JEROME A. MARK, U.S. Bureau of Labor Statistics
Papers: WILLIAM H. WALDORF, State University of New York-Binghamton, KENT KUNZE, LARRY ROSENBLUM, U.S. Bureau of Labor Statistics, AND MICHAEL TANNEN, University of the District of Columbia
New Measures of the Contribution of Education and Experience to U.S. Productivity Growth, 1948–85
WILLIAM GULLICKSON AND MICHAEL HARPER, U.S. Bureau of Labor Statistics
The Effects of Technology—Intensive Inputs on Productivity in Manufacturing Industries
LEO SVEIKAUSKAS, U.S. Bureau of Labor Statistics
Research and Development and Productivity Growth
Discussants: JACOB MINCER, Columbia University
M. ISHAQ NADIRI, New York University

8:00 A.M. THE ECONOMICS OF SCIENCE
Presiding: ARTHUR M. DIAMOND, JR., Ohio State University
Papers: SHARON G. LEVIN, University of Missouri-St. Louis, AND PAULA E. STEPHAN, Georgia State University
Obsolescence and Scientific Productivity
ARTHUR M. DIAMOND, JR., Ohio State University
The Determinants of a Scientist's Choice of Research Project
SUSAN FEIGENBAUM, DAVID LEVY, AND GORDON TULLOCK, George Mason University
The Economics of Econometric Research
Discussants: JOHN M. MCDOWELL, Arizona State University
MARIANNE A. FERBER, University of Illinois
W. LEE HANSEN, University of Wisconsin-Madison

8:00 A.M. LONG-RUN ECONOMIC GROWTH: WHAT HAVE WE LEARNED?
Presiding: (TO BE ANNOUNCED)
Papers: DALE JORGENSON, Harvard University
Technological Change and Economic Growth: An Evaluation of Recent Evidence
W. ERWIN DIEWERT, University of British Columbia, AND CATHY MORRISON, Tufts University
Total Factor Productivity in the United States, Japan, and Canada: New Measures and Estimates
MICHAEL J. BOSKIN, Stanford University
Taxation and Capital Formation: Lessons from the 1980s
Discussants: LAWRENCE H. SUMMERS, Harvard University
MARTIN N. BAILY, The Brookings Institution
PAUL DAVID, Stanford University

8:00 A.M. WHAT FUTURE FOR MONETARISM? (A Roundtable)
Presiding: ALLAN H. MELTZER, Carnegie-Mellon University
Panel: KARL BRUNNER, University of Rochester
ROBERT J. GORDON, Northwestern University
BENNETT MCCALLUM, Carnegie-Mellon University

8:00 A.M. AGING AND EMPLOYMENT
Presiding: ROBERT L. CLARK, North Carolina State University-Raleigh
Papers: MARJORIE HONIG AND CORDELIA W. REIMERS, Hunter College
Movers and Stayers Among the Elderly: Partial Retirement, Full Retirement, and Full-Time Work
DONALD O. PARSONS, Ohio State University
Industrial Differences in the Employment of the Aged
RANDALL O. FILER, Hunter College, AND PETER A. PETRI, Brandeis University
A Job-Characteristic Theory of Retirement
Discussants: RICHARD V. BURKHAUSER, Vanderbilt University
ROBERT M. HUTCHENS, Cornell University
JOSEPH F. QUINN, Boston College

8:00 A.M. SOCIAL AND ECONOMIC CONSEQUENCES OF THE INDUSTRIAL REVOLUTION
Presiding: ROBERT W. FOGEL, University of Chicago
Papers: GEORGE BOYER, Cornell University
Malthus Was Right After All
CLAYNE POPE, Brigham Young University
Mortality Cycles in Nineteenth-Century America
JEFFREY G. WILLIAMSON, Harvard University
Social and Economic Consequences of the Industrial Revolution
Discussants: ROBERT W. FOGEL, University of Chicago
MICHAEL HAINES, Wayne State University

8:00 A.M. ECONOMIC AND ETHICAL ISSUES IN ISLAMIC THOUGHT (Joint Session with the Association for Comparative Economic Studies)
Presiding: EBRAHIM HARRAF, Embry-Riddle University
Papers: TIMUR KURAN, University of Southern California
The Notion of Economic Justice in Contemporary Islamic Thought
STEVEN L. FLINT, University of Texas-Austin
Risk and Insurance in the Islamic Framework
L. DWIGHT ISRAELSON, Utah State University, AND RADI EL-BDOUR, Yarmouk University, Jordan
An Economic Analysis of Islamic Interest-Free Banking
Discussants: FREDERIC L. PRYOR, Swarthmore College
NADEEM U. HAQUE, World Bank

8:00 A.M. ANALYZING HEALTH OUTCOMES (Joint Session with the Health Economics Research Organization)
Presiding: WARREN GREENBERG, George Washington University
Papers: JOSEPH LIPSCOMB, JR., Duke University
Human Capital, Willingness to Pay, and Cost-Effectiveness Analyses of Screening for Birth Defects
ROBERT WOODWARD, Washington University
Cost Outcome Analysis of the Use of Intensive Care Units for Critically Ill Patients
MICHAEL GROSSMAN, City University of New York and National Bureau of Economic Research, AND THEODORE JOYCE, Iona College and National Bureau of Economic Research
Impact of Abortion on the Distribution of Birth Weight in New York City
Discussants: STUART SCHWEITZER, University of California-Los Angeles
BARBARA WOLFE, University of Wisconsin-Madison
HARRIET ORCUTT DULEEP, U.S. Commission on Civil Rights

8:00 A.M. THE HISPANIC ECONOMY IN THE UNITED STATES (Joint Session with the Hispanic Professors of Economics and Business)
Presiding: JORGE SALAZAR-CARRILLO, Florida International University and The Brookings Institution
Papers: GEORGE BORJAS, University of California-Santa Barbara
Hispanics in the Labor Market: A Comparison in the 1970 and 1980 Census Cross Section
ANTONIO JORGE, RAUL MONCARZ, IRMA TIRADO DE ALONSO, Florida International University, AND JORGE SALAZAR-CARRILLO, Florida International University and The Brookings Institution
The Migrant as a Businessman: The Impact of Hispanics on the Receiving South Florida Economy
FRANCISCO RIVERA-BATIZ, University of Pennsylvania
Shift in Labor Migration Patterns Between Puerto Rico and the United States: 1945 to 1985
GILBERT CARDENAS, Pan American University
The Economic Implications of Immigration Policy on Mexicans

Discussants: LUIS LOCAY, State University of New York-Stony Brook
MARTA TIENDA, University of Wisconsin-Madison
ROGER BETANCOURT, University of Maryland

8:00 A.M. CONFLICT AND PEACE ECONOMICS I: FACTOR ANALYSIS AND INTERNATIONAL SECURITY (Joint Session with the Peace Science Society International)
Presiding: (TO BE ANNOUNCED)
Papers: MURRAY WOLFSON, Oregon State University
Factor Analysis and the Dynamics of Arms Races
MARTIN MCGUIRE, University of Maryland
International Security, Factor Mobility, and Trade Control

8:00 A.M. POVERTY POLICY AND THE ECONOMICS OF POVERTY
Presiding: VICTOR FUCHS, Stanford University
Papers: ROBERT K. HAVEMAN, University of Wisconsin-Madison
The Evolution of Applied Economics: Legacies of the War on Poverty
TIMOTHY M. SMEEDING, University of Utah, AND BARBARA B. TORREY, U.S. Bureau of the Census
Comparative Analyses of Economic Well-Being and Poverty: Lessons from the Luxembourg Income Study
ISABEL SAWHILL, Urban Institute
Conceptualizing and Measuring the Underclass
Discussants: HENRY AARON, University of Maryland and The Brookings Institution
EUGENE SMOLENSKY, University of Wisconsin-Madison

10:15 A.M. STABILIZATION AND ECONOMIC GROWTH IN DEVELOPING COUNTRIES
Presiding: ANNE O. KRUEGER, World Bank
Papers: MOHSIN A. KHAN, World Bank
Stabilization and Economic Growth in Developing Countries: Some Policy Aspects
WILLEM H. BUITER, Yale University and National Bureau of Economic Research
Analytical Aspects of Stabilization and Structural Adjustment in Developing Countries
Discussants: JACOB A. FRENKEL, University of Chicago
MORRIS GOLDSTEIN, International Monetary Fund

10:15 A.M. LESSONS FROM HEALTH ECONOMICS
Presiding: VICTOR R. FUCHS, Stanford University
Papers: MARK V. PAULY, University of Pennsylvania
Nonprofit Firms in Medical Markets
VICTOR R. FUCHS, Stanford University, AND RICHARD ZECKHAUSER, Harvard University
Valuing Health and Other "Priceless" Commodities
JOSEPH P. NEWHOUSE, Rand Corporation
Econometrics in Health Economics
Discussants: BURTON WEISBROD, University of Wisconsin-Madison
CHARLES E. PHELPS, University of Rochester
ROGER D. FELDMAN, University of Minnesota

10:15 A.M. PHILOSOPHY AND ECONOMICS
Presiding: HAL R. VARIAN, University of Michigan
Papers: DANIEL M. HAUSMAN, Carnegie-Mellon University
Philosophy and Economics
MICHAEL S. MCPHERSON, Williams College and The Brookings Institution
Economics and Philosophy
Discussants: (TO BE ANNOUNCED)

10:15 A.M. MULTIPLE EQUILIBRIA IN MACROECONOMICS
Presiding: COSTAS AZARIADIS, University of Pennsylvania
Papers: PETER A. DIAMOND, Massachusetts Institute of Technology
Multiple Equilibria in Credit Markets
ROBERT J. SHILLER, Yale University
Ultimate Sources of Aggregate Variability
MICHAEL WOODFORD, University of Chicago
Self-Fulfilling Expectations and Business Cycles
Discussants: BEN S. BERNANKE, Princeton University
DOUGLAS W. DIAMOND, University of Chicago
ROBERT E. LUCAS, JR., University of Chicago

10:15 A.M. REGULATORY REFORM IN TRANSPORTATION
Presiding: PAUL W. MACAVOY, University of Rochester
Papers: JAMES M. MACDONALD, U.S. Department of Agriculture
Railroad Deregulation: Analyses of the Effects on Grain Transportation
NANCY L. ROSE, Massachusetts Institute of Technology
Union Wage Gains Under Regulation: Evidence from the Trucking Industry
KENNETH J. BUTTON, Loughborough University, England
The Impact of Transport Regulatory Reform in the United Kingdom
Discussants: PAUL W. MACAVOY, University of Rochester
PABLO T. SPILLER, Hoover Institution

10:15 A.M. FINANCE AND ECONOMICS
Presiding: MYRON S. SCHOLES, Stanford University
Papers: STEPHEN A. ROSS, Yale University
The Interrelations of Finance and Economics: Theoretical Perspectives
MICHAEL R. GIBBONS, Stanford University
The Interrelations of Finance and Economics: Empirical Perspectives
Discussants: WILLIAM BROCK, University of Wisconsin-Madison
MYRON S. SCHOLES, Stanford University
LAWRENCE H. SUMMERS, Harvard University

10:15 A.M. MULTINATIONALS AND TRADE IN SERVICES
Presiding: CHARLES P. KINDLEBERGER, Massachusetts Institute of Technology
Papers: ALAN M. RUGMAN, Dalhousie University
A Transaction Cost Approach to Trade in Services
HERBERT G. GRUBEL, Simon Fraser University
Trade in Services and the Multinational Enterprise
IRVING B. KRAVIS, University of Pennsylvania, AND ROBERT A. LIPSEY, National Bureau of Economic Research and Queens College
Production and Trade in Services by U.S. Multinationals
Discussants: RAYMOND VERNON, Harvard University
JAMES A. BRANDER, University of British Columbia
JAMES R. MARKUSEN, University of Western Ontario

10:15 A.M. ANTHROPOLOGY AND ECONOMICS
Presiding: GEORGE AKERLOF, University of California-Berkeley
Papers: LUIS LOCAY, State University of New York-Stony Brook
Fertility and Population Equilibrium in Primitive Economies
ROBERT TOWNSEND, University of Chicago
Contract Theory and the Medieval Manor: An Analysis of "Primitive" Economies
Discussants: DONALD N. MCCLOSKEY, University of Iowa
FREDERIC L. PRYOR, Swarthmore College
DANIEL GROSS, Hunter College

10:15 A.M. WOMEN IN COMPARATIVE PERSPECTIVE (Joint Session with the Association for Comparative Economic Studies)
Presiding: FRANCINE BLAU, University of Illinois-Urbana
Papers: ELIZABETH CLAYTON, University of Missouri-St. Louis
Age-Earnings Profiles of Women in the USSR
ROBERT GREGORY, Australian National University
Introducing Equal Pay and Comparable Worth: Implications and Consequences in Three Countries
LYNN TURGEON, Hofstra University
The Effects of the New Economic Mechanism on the Relative Position of Hungarian Women
ALENA HEITLINGER, Trent University
Variations in Population Policies: East-West Comparisons
PENELOPE PRIME, Emory University
A Comparison of Female Employment Patterns in Chinese Development Strategies: The Cases of Taiwan and Jiangsu Province
Discussants: PAUL GREGORY, University of Houston
MARIANNE A. FERBER, University of Illinois-Urbana

10:15 A.M. ODE GRADUATE STUDENT PAPER SESSION (Joint Session with Omicron Delta Epsilon)
Presiding: ALEX J. KONDONASSIS, University of Oklahoma
Papers: (TO BE ANNOUNCED)

10:15 A.M. DEVELOPMENT ECONOMISTS LOOK AT U.S. POLICY (Joint Session with the Society for Policy Modeling)
Presiding: IRMA ADELMAN, University of California-Berkeley
Papers: BELA BALASSA, World Bank and Johns Hopkins University
U.S. Trade Strategy
STANLEY LAWSON, St. John's University, AND BARRY HERMAN, United Nations
Does the United States Have a Foreign Debt Problem?
SHERMAN ROBINSON, University of California-Berkeley
Techniques of Policy Analysis
Discussants: GEORGE FEIWEL, University of Tennessee-Knoxville
BRIAN WRIGHT, University of California-Berkeley
ALBERT FISHLOW, University of California-Berkeley

12:30 P.M. AEA/AFA JOINT LUNCHEON
Presiding: GARY S. BECKER, University of Chicago
Speaker: GEORGE J. STIGLER, University of Chicago

2:30 P.M. WHITHER TAX REFORM?
Presiding: JOSEPH PECHMAN, The Brookings Institution
Panel: HENRY J. AARON, The Brookings Institution
MICHAEL J. BOSKIN, Stanford University
JAMES M. POTERBA, Massachusetts Institute of Technology
JOEL B. SLEMROD, University of Minnesota

2:30 P.M. IN SEARCH OF HISTORICAL ECONOMIES: MARKET BEHAVIOR IN PAST TIMES
Presiding: DOUGLASS C. NORTH, Washington University-St. Louis
Papers: DONALD N. MCCLOSKEY, University of Iowa and University of York
The Economy of Village Agricultфre in the Middle Ages
DAVID W. GALENSON, University of Chicago
Market Behavior in Colonial America
LANCE E. DAVIS, California Institute of Technology
Government Policy, Profit Opportunities, and Economic Reality: The Political Economy of the British Raj
JAMES J. HECKMAN, University of Chicago
Racial Discrimination and Industrialization: South Carolina, 1910–70
Discussants: THEODORE W. SCHULTZ, University of Chicago
DOUGLASS C. NORTH, Washington University-St. Louis

2:30 P.M. DO PEOPLE BEHAVE RATIONALLY?
Presiding: ROBERT H. FRANK, Cornell University
Papers: HAL R. VARIAN, University of Michigan
Testing for Rational Choice
ROBERT H. FRANK, Cornell University
The Strategic Role of the Emotions
GORDON WINSTON, Williams College
How Rational is it Rational to Be? The Timing of its Payoffs and the Role of Time Preferences
LAURENCE R. IANNACCONE, University of Santa Clara
An Economic Model of Religious Participation
Discussants: GARY S. BECKER, University of Chicago
GEORGE A. AKERLOF, University of California-Berkeley

2:30 P.M. THE FAMILY
Presiding: WARREN SANDERSON, State University of New York-Stony Brook
Paper: ROBERT WILLIS, University of Chicago
What Have We Learned from the Economics of the Family?
Discussants: KENNETH WOLPIN, Ohio State University
ROBERT POLLAK, University of Pennsylvania
T. PAUL SCHULTZ, Yale University

2:30 P.M. ANALYSIS OF POVERTY IN DEVELOPING COUNTRIES
Presiding: ROBERT E. EVENSON, Yale University
Papers: GIAN S. SAHOTA, Vanderbilt University
A General Theory of Poverty and Income Distribution and Applications
SHERWIN ROSEN, University of Chicago
Family Structure and Poverty in Panama
Discussants: IRMA ADELMAN, University of California-Berkeley
GARY S. FIELDS, Cornell University
CLIVE BELL, Johns Hopkins University

2:30 P.M. ECONOMICS AND POLITICS
Presiding: GEORGE J. STIGLER, University of Chicago
Papers: SAM PELTZMAN, University of Chicago
Economic Conditions and Gubernatorial Elections
THOMAS ROMER AND HOWARD ROSENTHAL, Carnegie-Mellon University
Agenda Control and Public Finance
JOHN FEREJOHN, Hoover Institution
The Theory of Legislative Bargaining
Discussants: ASSAR LINDBECK, Institute for International Economic Studies, Stockholm
JAMES SNYDER, University of Chicago

2:30 P.M. REDISTRIBUTIONAL ASPECTS OF CONSTITUTIONAL REFORMS IN THE MONETARY SYSTEM (Joint Session with the Association for the Study of the Grants Economy)
Presiding: KENNETH E. BOULDING, University of Colorado-Boulder
Papers: JOHN H. HOTSON, University of Waterloo
The Case for Financial Reform: Or, Was Irving Fisher Right After All?
MARTIN H. WOLFSON, Board of Governors of the Federal Reserve System
Institutional Change and the Financial System
MARVIN S. GOODFRIEND, Federal Reserve Bank of Richmond
Distributional Aspects of Monetary Policy
Discussants: GORDON TULLOCK, George Mason University
RICHARD D. BARTEL, *Challenge* Magazine
THOMAS F. WILSON, American Fletcher National Bank

2:30 P.M. PRICE VARIABILITY IN NATURAL RESOURCE MARKETS: CAUSES AND CONSEQUENCES (Joint Session with the Association of Environmental and Resource Economics)
Presiding: ERNST BERNDT, Massachusetts Institute of Technology
Papers: JAMES SWEENEY, Stanford University
Price Instability in Natural Resource Markets
ROBERT PINDYCK, Massachusetts Institute of Technology, AND SAMAN MAJD, University of Pennsylvania
Modularity, Learning and Sequential Investment in Natural Resources
WILLIAM HOGAN AND PAUL LEIBY, Harvard University
Oil Market Risk Analysis
Discussants: (TO BE ANNOUNCED)

2:30 P.M. CONFLICT AND PEACE ECONOMICS II: PANEL ON MILITARY EXPENDITURES AND NATIONAL SECURITY (Joint Session with the Peace Science Society International)
Presiding: (TO BE ANNOUNCED)
Panel: MANCUR OLSON, University of Maryland
MARTIN J. BAILEY, U.S. Department of State
BARRY NALEBUFF, Princeton University
WALTER ISARD, Cornell University

2:30 P.M. HOW DO IMMIGRANTS DO IN THE U.S. LABOR MARKET?
Presiding: FINIS WELCH, University of California-Los Angeles
Papers: GEORGE J. BORJAS, University of California-Santa Barbara
The Earnings of Immigrants: Perspectives in Recent Research
BARRY R. CHISWICK, University of Illinois-Chicago
Further Thoughts on the Labor Market Adjustment of Immigrants
Discussants: CORDELIA REIMERS, Hunter College
ROBERT TOPEL, University of California-Los Angeles

2:30 P.M. CONTINUING BLACK POVERTY (Joint Session with the National Economic Association)
Presiding: DAVID H. SWINTON, Clark College
Papers: SHELDON DANZINGER, University of Wisconsin, AND PETER GOTTSHALK, Bowdoin College
Earnings Inequaltiy and Poverty, 1970 to 1979
WILLIAM WILSON, University of Chicago
Joblessness and Poverty
WILLIAM A. DARITY, JR., University of North Carolina-Chapel Hill, AND SAMUEL MYERS, JR., University of Pittsburgh
Do Transfer Payments Keep the Poor in Poverty?
DAVID H. SWINTON, Clark College
Economic Theory and Poverty
Discussant: BARBARA A. P. JONES, Clark College

8:00 P.M. RICHARD T. ELY LECTURE
Presiding: GARY S. BECKER, University of Chicago
Paper: JUDGE RICHARD A. POSNER, United States Court of Appeals for the Seventh District
The Law and Economics Movement

Monday, December 29, 1986

8:00 A.M. STUDIES OF THE CHINESE ECONOMY
Presiding: NICHOLAS R. LARDY, University of Washington
Papers: BARRY J. NAUGHTON, University of Oregon
Real Urban Income in China: Prices and Rations
CHRISTINE WONG, Mount Holyoke College
Maoism and Development: Rural Industrialization and Intersectoral Resource Flows During the Cultural Revolution Decade
Discussants: JOHN FEI, Yale University
THOMAS P. LYONS, Dartmouth College
TERRY SICULAR, Stanford University

8:00 A.M. EDUCATION AND ECONOMICS
Presiding: ROBERT T. MICHAEL, University of Chicago
Papers: JAMES S. COLEMAN, University of Chicago
Social Capital in the Creation of Human Capital
JERE BEHRMAN, ROBERT POLLAK, AND PAUL J. TAUBMAN, University of Pennsylvania
Inter- and Intra-Generational Determinants of Education
Discussants: RICHARD J. MURNANE, Harvard University
JOHN M. MCDOWELL, Arizona State University

8:00 A.M. CONTRIBUTED PAPERS ON INDUSTRIAL ORGANIZATION: THEORETICAL
Presiding: STEVEN C. SALOP, Georgetown University
Papers: MARK BAGNOLI, University of Michigan
Noncollusive Incentives to Form Trade Associations
FRANKLIN ALLEN AND GERALD R. FAULHABER, University of Pennsylvania
Reputation and Quality Noise
LEWIS T. EVANS, Victoria University of Wellington, New Zealand, AND STEVEN G. GARBER, Carnegie-Mellon University
A Theory of Constraints that Benefit Rational Public-Utility Regulators
Discussants: CARL SHAPIRO, Princeton University
MICHAEL D. WHINSTON, Harvard University

8:00 A.M. BOUNDED RATIONALITY AND COOPERATION
Presiding: THEODORE GROVES, University of California-San Diego
Papers: AVRAM NYMAN, Hebrew University and State University of New York-Stony Brook
Bounded Rationality and Cooperation
SYLVAIN SORAIN, University of Strasbourg, France, AND ROBERT AUMANN, Math Science Research Institute
Bounded Rationality and Cooperation
Discussants: ERIC MASKIN, Harvard University
C. BARTLETT MCGUIRE, University of California-Berkeley
JOHN ROBERTS, Stanford University

8:00 A.M. EMPIRICAL STUDIES IN BANKING AND CREDIT
Presiding: PAUL WACHTEL, New York University
Papers: ARIE MELNIK, Haifa University, Israel, GEORGE SOFIANOS, AND PAUL WACHTEL, New York University
Credit Lines and the Transmission of Monetary Policy
JOSEPH G. HAUBRICK, University of Pennsylvania
Money Supply, Bank Failures and Credit Crunches: Lesson from the Great Depression in Canada
ERNST JUERG WEBER, California State University-Northridge
Parallel Currencies in Switzerland, 1826–50
Discussants: STEPHEN KING, Stanford University and Federal Reserve Bank of New York
EUGENE WHITE, Rutgers University
WARREN WEBER, Federal Reserve Bank of Minneapolis

8:00 A.M. DYNAMIC MODELS OF TAXATION
Presiding: DAVID F. BRADFORD, Princeton University
Papers: KENNETH L. JUDD, Northwestern University
A Dynamic Theory of Factor Taxation
LAURENCE J. KOTLIKOFF, Boston University, AND ALAN J. AUERBACH, University of Pennsylvania
Evaluating Fiscal Policy with a Dynamic Simulation Model
Discussants: JOHN B. SHOVEN, Stanford University
JAMES M. POTERBA, Massachusetts Institute of Technology
B. DOUGLAS BERNHEIM, Stanford University

8:00 A.M. RULES VS. DISCRETION IN MACROECONOMIC POLICY
Presiding: BENNETT T. MCCALLUM, Carnegie-Mellon University
Papers: ALBERTO ALESINA, Carnegie-Mellon University
Rules, Discretion and Political Cycles
KENNETH S. ROGOFF, University of Wisconsin-Madison
Policy Credibility in a Multi-Sovereign Environment
ROBERT STAIGER, Stanford University, AND GUIDO TABELLINI, University of California-Los Angeles
On the Time-Consistency of Optimal Redistributive Fiscal Policies
Discussants: EDWARD GREEN, University of Pittsburgh
ROBERT J. BARRO, University of Rochester
N. GREGORY MANKIW, Harvard University

8:00 A.M. A SURVEY OF HOUSING ECONOMICS (Joint Session with the American Real Estate and Urban Economic Association)
Presiding: JOHN C. WEICHER, American Enterprise Institute
Papers: RICHARD F. MUTH, Emory University
Housing Market Dynamics
RICHARD J. ARNOTT, Queens University
Recent Theoretical Developments in the Microeconomics of Housing
EDGAR O. OLSEN, University of Virginia
The Demand and Supply of Housing Services: A Critical Survey of the Empirical Literature
Discussants: JOHN M. QUIGLEY, University of California-Berkeley
WILLIAM C. WHEATON, Massachusetts Institute of Technology
GEORGE F. TOLLEY, University of Chicago

8:00 A.M. LONG-RUN INDUSTRIALIZATION: SELECTED ASIAN ECONOMIES (Joint Session with the Committee on Asian Economic Studies)
Presiding: RICHARD KOSOBUD, University of Illinois-Chicago
Papers: RICHARD HOOLEY, University of Pittsburgh
Industrial Productivity Growth in Southeast Asia
CHARLES PEARSON AND JAMES RIEDEL, Johns Hopkins University
Exchange Rate Protection: Case Study of Japan
MANORANJAN DUTTA, Rutgers University
Long-Run Industrialization of Selected Asian Economies: Impact on Structural Changes
Discussants: ROMEO M. BAUTISTA, International Food Research Institute and University of the Philippines, Manila
AKIRA TAKAYAMA, Southern Illinois University

8:00 A.M. ODE CHAPTER ADVISERS AND REGIONAL DIRECTORS SESSION (Joint Session with Omicron Delta Epsilon)
Presiding: G. RANDOLPH RICE, Louisiana State University
Papers: (TO BE ANNOUNCED)

10:15 A.M. ROUNDTABLE ON TEACHING UNDERGRADUATE COURSES IN QUANTITATIVE METHODS
Presiding: W. LEE HANSEN, University of Wisconsin-Madison
Papers: GLEN G. CAIN, University of Wisconsin-Madison
The Centrality of Economics in Teaching Quantitative Methods in Economics
WILLIAM E. BECKER, JR., Indiana University
Teaching Quantitative Methods to Undergraduate Economics Students
VIJAYA G. DUGGAL, Widener University
Coping With the Diversity of Student Aptitudes and Interests in Quantitative Methods Courses

10:15 A.M. REFORMING THE INTERNATIONAL MONETARY SYSTEM
Presiding: JACQUES I. POLAK, International Monetary Fund
Papers: PETER B. KENEN, Princeton University
Exchange Rate Management: What Role for Intervention?
JOHN WILLIAMSON, Institute for International Economics
Exchange Rate Management: The Role of Target Zones
JACOB A. FRENKEL, University of Chicago
The International Monetary System in Need for Repair: Will Fix Exchange Rates Do the Job?
Discussants: ALLAN H. MELTZER, Carnegie-Mellon University
MAURICE OBSTFELD, Columbia University

10:15 A.M. THE ECONOMIC ANALYSIS OF TAXPAYER COMPLIANCE
Presiding: JAMES M. POTERBA, Massachusetts Institute of Technology
Papers: SUZANNE SCOTCHMER, University of California-Berkeley
Equity and Tax Enforcement
JAMES M. POTERBA, Massachusetts Institute of Technology
Tax Evasion and Effective Tax Rates
JEFFREY A. DUBIN, LOUIS L. WILDE, California Institute of Technology, AND MICHAEL GRAETZ, Yale Law School
New Econometric Evidence on the Determinants of Taxpayer Compliance
Discussants: JENNIFER F. REINGANUM, California Institute of Technology
JAMES HINES, Princeton University
JOEL B. SLEMROD, University of Minnesota

10:15 A.M. CRIME, PUNISHMENT, AND THE BUSINESS CYCLE
Presiding: SAMUEL L. MYERS, JR., University of Pennsylvania
Papers: RICHARD FREEMAN, Harvard University
Joblessness and Crime Among Inner-City Youth: Analysis of the NBER *Survey of Inner-City Youth*
DARIO MELOSSI, University of California-Santa Barbara
Political Business Cycle and Imprisonment Rates: The Case of Italy, 1896–1965
SAMUEL L. MYERS, JR. AND WILLIAM SABOL, University of Pennsylvania
Racial Differences in Incarceration and the Business Cycle
HAROLD L. VOTEY, JR. AND LLAD PHILLIPS, University of California-Santa Barbara
Economic Opportunity, Family Support, and Crimes by Youth
Discussants: RICHARD M. MCGAHEY, New York University
LLAD PHILLIPS, University of California-Santa Barbara

10:15 A.M. THE INTERNATIONAL DEBT CRISIS
Presiding: THOMAS P. WILLETT, Claremont Graduate School and Claremont Men's College
Papers: ANNE O. KRUEGER, World Bank
Debt, Capital Flows, and *LDC* Growth
STANLEY FISCHER, Massachusetts Institute of Technology
Resolving the International Debt and Growth Crisis
H. ROBERT HELLER, Bank of America
The Debt Crisis and the Future of International Bank Lending
Discussants: MANUEL GUITIAN, International Monetary Fund
ARMEANE CHOKSI, World Bank

10:15 A.M. NEW ADVANCES IN INPUT-OUTPUT ANALYSIS: SESSION IN HONOR OF WASSILY LEONTIEF
Presiding: ANNE CARTER, Brandeis University
Papers: LAWRENCE KLEIN, University of Pennsylvania
Econometric Aspects of Input-Output Analysis
KAREN R. POLENSKE, Massachusetts Institute of Technology
New Perspectives on Input-Output Analysis: A Cross-Country Comparison
FAYE DUCHIN, New York University
Closure of a Dynamic Input-Output Model to Investment and Consumption
ADAM ROSE, West Virginia University
Input-Output Analysis of Income Distribution
Discussants: CLOPPER ALMON, JR., University of Maryland
SHERMAN ROBINSON, University of California-Berkeley

10:15 A.M. SYMPOSIUM ON THE FUTURE OF U.S. ANTITRUST POLICY
Presiding: RICHARD L. SCHMALENSEE, Massachusetts Institute of Technology
Panel: RICHARD A. POSNER, Seventh Circuit Court of Appeals and University of Chicago
DONALD F. TURNER, Kirkland & Ellis
STEVEN C. SALOP, Georgetown University
LAWRENCE J. WHITE, New York University

10:15 A.M. THE CONTRIBUTION OF KEYNES AFTER 50 YEARS
Presiding: ROBERT J. GORDON, Northwestern University
Papers: WILLEM BUITER, Yale University
Market Failure, Political Failure, and Stabilization Policy
ALAN BLINDER, Princeton University
Keynes After Lucas
GEORGE AKERLOF AND JANET YELLEN, University of California-Berkeley
Rational Models of "Irrational" Behavior
Discussants: BENNETT T. MCCALLUM, Carnegie-Mellon University
ROBERT J. BARRO, University of Rochester
ROBERT E. LUCAS, JR., University of Chicago

10:15 A.M. ECONOMISTS VIEW ON GRANTS ECONOMICS: A CONVERGENCE HYPOTHESIS? (Joint Session with the Association for the Study of the Grants Economy)
Presiding: ABRAM BERGSON, Harvard University
Paper: ALAN A. BROWN, University of Windsor, JANOS HORVATH, Butler University, AND EGON NEUBERGER, State University of New York-Stony Brook
Grants Economics: Theory and Applications
Roundtable Discussion: ABRAM BERGSON, Harvard University
KENNETH E. BOULDING, University of Colorado-Boulder
JOSEPH A. PECHMAN, The Brookings Institution
ROBERT M. SOLOW, Massachusetts Institute of Technology
MURRAY L. WEIDENBAUM, Washington University
BURTON A. WEISBROD, University of Wisconsin-Madison

10:15 A.M. ANALYSES OF THE IMPACTS OF PROSPECTIVE PAYMENTS SYSTEMS (Joint Session with the Health Economics Research Organization)
Presiding: DONALD E. YETT, University of Southern California
Papers: JUDITH FEDER, JACK HADLEY, Georgetown University, AND STEPHEN ZUCKERMAN, Urban Institute
Selected Analyses of *PPS*'s Impact on Hospital Behavior
WILLIAM CUSTER, JAMES MOSER, ROBERT MUSACCHIO, AND RICHARD WILLKE, American Medical Association
Impact of Medicare's Prospective Payment System on Hospital Attributes, Performance, and Quality of Care
FRANK SLOAN, Vanderbilt University, MICHAEL MORRISEY, University of Alabama-Birmingham, AND JOSEPH VALVONA, Vanderbilt University
Hospital Case Mix Variation in Response to Medicare Prospective Payment
Discussants: RICHARD ARNOULD, University of Illinois
STUART GUTERMAN, Health Care Financing Administration
RONALD VOGEL, University of Arizona

10:15 A.M. ECONOMETRIC RESEARCH AND THE DEVELOPMENT OF ECONOMIC ANALYSIS (Joint Session with the History of Economics Society)
Presiding: A. W. COATS, Duke University and University of Nottingham, England
Papers: NANCY WULWICK, Le Moyne College
The Bewitching Phillips Curve
BASIL MOORE, Wesleyan College
An Historical Perspective to the Econometrics of Money and Income
INGRID RIMA, Temple University
Econometric Research and Changing Perspectives of Unemployment
Discussants: JOHN SMITHIN, York University
WILLIAM BUTOS, Trinity College
CORNELIS LOS, Federal Reserve Bank of New York

10:15 A.M. THE ECONOMICS OF DISCRIMINATION THIRTY YEARS LATER
Presiding: ORLEY C. ASHENFELTER, Princeton University
Papers: THOMAS D'AMICO, St. Peter's College
Economic Theories of Discrimination
FRANCINE D. BLAU AND MARIANNE A. FERBER, University of Illinois-Urbana
Discrimination: Empirical Evidence from the United States
ORLEY C. ASHENFELTER, Princeton University, AND RONALD L. OAXACA, University of Arizona
Discrimination: Empirical Evidence from Abroad
Discussants: GARY S. BECKER, University of Chicago
ISABEL V. SAWHILL, Urban Institute

12:30 P.M. LUNCHEON IN HONOR OF NOBEL LAUREATE, FRANCO MODIGLIANI
Presiding: GARY S. BECKER, University of Chicago
Speaker: ROBERT MERTON, Massachusetts Institute of Technology

2:30 P.M. GENDER DIFFERENCES IN BEHAVIOR AT HOME AND AT WORK
Presiding: REBECCA BLANK, Princeton University
Papers: WILLIAM S. CUSTER, American Medical Association, AND DENISE DIMON, University of San Diego
The Influence of Gender in the Choice of Physicians' Practice Mode
DAVID E. BLOOM AND SANDERS KORENMAN, Harvard University
Gender Differences in Consumer Spending
JANICE MADDEN, University of Pennsylvania
Gender Differences in the Cost of Displacement
FAITH ANDO, Jaca Corporation
Sex Discrimination in Commercial Bank Credit
Discussants: REBECCA BLANK, Princeton University
JANET KOHLHASE, University of Houston

2:30 P.M. DEVELOPMENT STUDIES OF THE INDIAN ECONOMY
Presiding: EDWIN S. MILLS, Princeton University
Papers: CHARLES M. BECKER, Vanderbilt University, EDWIN S. MILLS, Princeton University, AND JEFFREY G. WILLIAMSON, Harvard University
Forces Underlying Indian Urbanization and Economic Growth, 1960–80
T. N. SRINIVASAN, Yale University
Rural Credit Markets: Segmentation, Rationing and Spillover
Discussants: SHERMAN ROBINSON, University of California-Berkeley
LARRY E. WESTPHAL, Swarthmore College

2:30 P.M. GAME THEORY AND INDUSTRIAL ORGANIZATION
Presiding: ROBERT D. WILLIG, Princeton University
Papers: DREW D. FUDENBERG, University of California-Berkeley, AND JEAN TIROLE, Massachusetts Institute of Technology
Dynamic Market Competition
JOHN ROBERTS, Stanford University
Informational Asymmetries and Industrial Competition
Discussants: TIMOTHY F. BRESNAHAN, Stanford University
RICHARD L. SCHMALENSEE, Massachusetts Institute of Technology
ROBERT D. WILLIG, Princeton University

2:30 P.M. BUDGET DEFICITS AND THE ECONOMY
Presiding: PAUL MACAVOY, University of Rochester
Panel: ROBERT J. BARRO, University of Rochester
MANUEL JOHNSON, Board of Governors of the Federal Reserve System
FRANCO MODIGLIANI, Massachusetts Institute of Technology
LAWRENCE H. SUMMERS, Harvard University

2:30 P.M. THE NEW POLITICAL ECONOMIC HISTORY
Presiding: THEODORE W. SCHULTZ, University of Chicago
Papers: DOUGLASS C. NORTH AND BARRY R. WEINGAST, Washington University-St. Louis
Political Structure and Economic Performance in Early Modern Europe
PHILLIP HOFFMAN, California Institute of Technology
The Political Obstacles to the Growth of French Agriculture in the Eighteenth Century
JOHN J. WALLIS, University of Maryland
The Political Economy of New Deal Fiscal Federalism
Discussants: MANCUR OLSON, University of Maryland
ROBERT BATES, Duke University

2:30 P.M. THE BEHAVIOR OF EXCHANGE RATES
Presiding: JACOB A. FRENKEL, University of Chicago
Papers: RUDIGER W. DORNBUSCH, Massachusetts Institute of Technology
Exchange Rate Movements and Relative Prices
MICHAEL MUSSA, Council of Economic Advisers
The Exchange Rate Regime and the Real Exchange Rate
PETER ISARD, International Monetary Fund
Lessons from Empirical Models of Exchange Rates
Discussants: ROBERT J. HODRICK, Northwestern University
ALAN C. STOCKMAN, University of Rochester

2:30 P.M. THE MANAGERIAL LABOR MARKET (Joint Session with the Association of Managerial Economists)
Presiding: MARK HIRSCHEY, Rice University
Papers: GEORGE P. BAKER, Harvard University
Compensation in Hierarchies
ANUP AGRAWAL AND GERSHON N. MANDELKER, University of Pittsburgh
Management Compensation and Corporate Capital Expenditures
MICHAEL C. JENSEN, Harvard University and University of Rochester, AND KEVIN J. MURPHY, University of Rochester
Are Executives Underpaid?
Discussants: EDWARD LAZEAR, University of Chicago and Hoover Institution
SHERWIN ROSEN, University of Chicago

2:30 P.M. THE VANISHING BLACK FAMILY (Joint Session with the National Economic Association)
Presiding: SAMUEL L. MYERS, JR., University of Pittsburgh
Papers: WILLIAM BRADFORD, University of Maryland
Changing Family Structure and Wealth Among Black Families
WILLIAM DARITY, JR., University of North Carolina-Chapel Hill, AND SAMUEL L. MYERS, JR., University of Pittsburgh
The Marginalization of Black Males and the Rise of Female-Headed Families
GREG DUNCAN, University of Michigan
Welfare and the Marriage Decision of Black Women
Discussants: GLENN LOURY, Harvard University
ISABEL SAWHILL, Urban Institute

2:30 P.M. INDUSTRIAL POLICIES, EXCHANGE RATES, AND PROTECTIONISM (Joint Session with the Society for Policy Modeling)
Presiding: DOMINICK SALVATORE, Fordham University
Papers: JAGDISH N. BHAGWATI, Columbia University
Protectionism: Old Wine in New Bottles
PAUL R. KRUGMAN, Massachusetts Institute of Technology
Targeted Industrial Policies: Theory and Evidence
RONALD I. MCKINNON, Stanford University
Objectives for International Negotiations for Stabilizing Exchange Rates
Discussants: JOHN F. O. BILSON, University of Chicago
ROBERT W. STAIGER, Stanford University

2:30 P.M. 1986 DISTINGUISHED LECTURE ON ECONOMICS IN GOVERNMENT (Joint Session with the Society of Government Economists)
Presiding: DOUGLAS L. ADKINS, Adkins Associates, Inc.
Introduction: GARY S. BECKER, University of Chicago
Speaker: JOSEPH A. PECHMAN, The Brookings Institution
Tax Reform: Theory and Practice

4:30 P.M. PRESIDENTIAL ADDRESS AND BUSINESS MEETING
Presiding: GARY S. BECKER, University of Chicago
Speaker: ALICE M. RIVLIN, The Brookings Institution
Economics and the Political Process

Tuesday, December 30, 1986

8:00 A.M. EMPIRICAL IMPLICATION OF NONLINEAR DYNAMICS: GENERALIZED TESTS FOR WHITENESS, INDEPENDENCE, AND FORECAST ABILITY
Presiding: MICHAEL ROTHSCHILD, University of California-San Diego
Papers: WILLIAM BROCK, University of Wisconsin-Madison
Empirical Implication of Nonlinear Dynamics: Generalized Tests for Whiteness, Independence and Forecast Ability
Discussants: MICHAEL WOODFORD, University of Chicago
LARS PETER HANSEN, University of Chicago
CHRISTOPHER SIMMS, University of Massachusetts

8:00 A.M. REAL BUSINESS CYCLE THEORY: WHAT DOES IT EXPLAIN?
Presiding: CHARLES I. PLOSSER, University of Rochester
Papers: MARTIN EICHENBAUM AND KENNETH J. SINGLETON, Carnegie-Mellon University
Money and Real Business Cycles
STEVEN DAVIS, University of Chicago
Sectoral Shifts and Real Business Cycles
JOHN LONG AND CHARLES I. PLOSSER, University of Rochester
Relative Prices, Sectoral Outputs and the Business Cycle
Discussant: ROBERT G. KING, University of Rochester

8:00 A.M. EMERGENT MARKETS FOR COMMON PROPERTY RESOURCES
Presiding: ALEX G. VICAS, McGill University
Papers: ANTHONY D. SCOTT, University of British Columbia
The Emergence of Marketable Fishery Licenses
ROBERT W. HAHN AND GORDON L. HESTER, Carnegie-Mellon University
Implementing "Radical" Ideas Incrementally: A Case Study of EPA's Emission Trading Program
MICHAEL E. LEVINE, University of Southern California
Markets in Airport Landing Rights
HARVEY L. LEVIN, Hofstra University
Emergent Markets for Orbit Spectrum Assignments—An Idea Whose Time Has Come
Discussants: CLAS G. WIHLBORG, Claremont Graduate School
SEVERIN BORENSTEIN, University of Michigan

8:00 A.M. DETERRENCE AND ENFORCEMENT OF LAWS
Presiding: MICHAEL K. BLOCK, U.S. Sentencing Commission
Papers: ISAAC EHRLICH, State University of New York-Buffalo
The Economics of Crime: Theory and Evidence
STEVEN SHAVELL, Harvard University
Deterrence vs. Incapacitation
Discussants: MICHAEL K. BLOCK, U.S. Sentencing Commission
PHILIP J. COOK, Duke University
DAVID D. FRIEDMAN, Tulane University

8:00 A.M. NEW THEORIES OF ECONOMIC GROWTH
Presiding: M. MAJUMDAR, Cornell University
Papers: ROBERT E. LUCAS, JR., University of Chicago
On the Mechanics of Economic Development

PAUL ROMER, University of Rochester
Growth Based on Increasing Returns Due to Specialization
EDWARD C. PRESCOTT, University of Minnesota
Organizational Arrangements and Growth

8:00 A.M. RETHINKING THE THEORY OF INTERNATIONAL TRADE
Presiding: RONALD W. JONES, University of Rochester
Papers: WILFRED J. ETHIER, University of Pennsylvania
Revising the Basic Trade Model
JAMES R. MARKUSEN, University of Western Ontario
The New Theory of Multinational Enterprise
PAUL R. KRUGMAN, Massachusetts Institute of Technology
Industrial Organization and Trade Theory: What Have We Learned?
Discussants: JAGDISH BHAGWATI, Columbia University
AVINASH DIXIT, Princeton University

8:00 A.M. MACROECONOMIC BEHAVIOR IN SELECTED ASIAN COUNTRIES: THE PEOPLE'S REPUBLIC OF CHINA, INDIA, AND JAPAN (Joint Session with the Committee on Asian Economic Studies)
Presiding: LAWRENCE KLEIN, University of Pennsylvania
Papers: JAMES T. H. TSAO, U.S. International Trade Commission and Georgetown University
Chinese Foreign Trade and Devaluation of the Yuan
K. KRISHNAMURTY, P. D. SHARMA, AND H. KRISHNASWAMY, Institute of Economic Growth, Delhi, India
Determinants of Savings Rate in India: A Quantitative Analysis
RALPH TRYON AND R. SHAN CRAIG, Federal Reserve Board
An Empirical Analysis of Japanese Trade in Goods and Services
Discussants: BELA BALASSA, Johns Hopkins University and World Bank
T. N. SRINIVASAN, Yale University
YUAN-LI WU, Hoover Institution and University of San Francisco

8:00 A.M. REFLECTIONS ON THE ECONOMICS OF THE SWEDISH SCHOOL (Joint Session with the History of Economics Society)
Presiding: DONALD WALKER, Indiana University of Pennsylvania
Papers: LARS JONUNG, University of Lund, Sweden
Why the Swedish School Failed
HANS BREMS, University of Illinois-Urbana
(Title to be announced)
LARRY SAMUELSON, Pennsylvania State University
The Wicksell Effect: Some Recent Treatments
Discussants: JAMES YOHE, Duke University
MEIR KOHN, Dartmouth College
INGRID RIMA, Temple University

8:00 A.M. THE IMPACT OF DEREGULATION ON EFFICIENCY IN ENERGY, TELECOMMUNICATIONS, AND TRANSPORT (Joint Session with the Transportation and Public Utilities Group)
Presiding: FREDERICK J. BEIER, University of Minnesota
Papers: MILTON Z. KAFOGLIS, Emory University
New Problems in a Deregulated Environment
MILTON RUSSELL, U.S. Environmental Protection Agency
Environmental Protection: A New Era of Regulation
DON H. PICKRELL, EDWARD L. RAMSDELL, AND RICHARD J. HORN, U.S. Department of Transportation-Transportation Systems Center
Changing Structure of the Air Carrier Industry Under Deregulation
Discussants: HARRY N. TREBING, Michigan State University
RICHARD GRITTA, University of Portland

8:00 A.M. CONTRIBUTED PAPERS ON INDUSTRIAL ORGANIZATION: EMPIRICAL
Presiding: LAWRENCE J. WHITE, New York University
Papers: KEITH J. CROCKER, University of Virginia, AND SCOTT E. MASTEN, University of Michigan
Unilateral Options and Contract Length in the Natural Gas Industry
IAN H. DOMOWITZ, R. GLENN HUBBARD, AND BRUCE C. PETERSEN, Northwestern University
Demand Fluctuations and Price Adjustment: A Panel-Data Study of U.S. Manufacturing
LINDA R. STANLEY, Oregon State University, AND JOHN T. TSCHIRHART, University of Wyoming
Consumer Response to Mandatory Information Disclosures: The Case of the Food Products Industry

Discussants: PAUL W. MACAVOY, University of Rochester
NANCY L. ROSE, Massachusetts Institute of Technology

10:15 A.M. FINANCIAL INTERMEDIATION AND ECONOMIC ACTIVITY
Presiding: LEE WAKEMAN, Chemical Bank of New York
Papers: DOUGLAS W. DIAMOND, University of Chicago
Reputations and Financial Intermediation
ROBERT G. KING, University of Rochester
Financial Intermediation and Economic Growth
ROBERT TOWNSEND, University of Chicago
Intermediation in Transactions Costs Economies
Discussants: JOSEPH G. HAUBRICH, University of Pennsylvania
JOHN BOYD, Federal Reserve Bank of Minneapolis
FISCHER BLACK, Goldman Sachs & Co.

10:15 A.M. GENDER AND BARGAINING POWER IN PUBLIC AND PRIVATE LABOR MARKETS
Presiding: ELYCE ROTELLA, Indiana University
Papers: SHULAMIT KHAN, University of California-Irvine
Economic Implications of Public Sector Comparable Worth: A Case Study of San Jose
DEB FIGART, American University
Gender, Unions, and Internal Labor Markets: Evidence from the Public Sector in Two States
DALLAS CULLEN, ALICE NAKAMURA, AND MASAO NAKAMURA, University of Alberta
Have Equal Opportunity/Affirmative Action Programs Had Any Impact on the Occupational Segregation of U.S. Women?
JAYNE DEAN, Wagner College
Sex Segregation and the Differential Bargaining Power of Workers
Discussants: ELYCE ROTELLA, Indiana University
ANDREA BELLER, University of Illinois

10:15 A.M. STOPPING HIGH INFLATION
Presiding: MANUEL GUITIAN, International Monetary Fund
Papers: STANLEY FISCHER, Massachusetts Institute of Technology
Stopping Inflation in Israel
JEFFREY SACHS, Harvard University
The Bolivian Stabilization
DANIEL HEYMAN, CEPAL
The Austral Plan
ELIANA A. CARDOSO AND RUDIGER DORNBUSCH, Massachusetts Institute of Technology
Brazil's Plan Tropical
Discussants: PETER GARBER, Brown University
GUILLERMO CALVO, University of Pennsylvania

10:15 A.M. ARBITRATION AND THE NEGOTIATION PROCESS
Presiding: HARRY C. KATZ, Cornell University
Papers: ORLEY ASHENFELTER, Princeton University
Arbitrator Behavior
MAX BAZERMAN, Northwestern University, AND HENRY FARBER, Massachusetts Institute of Technology
Decision Making in Negotiation and Arbitration
DAVID E. BLOOM, Harvard University
Negotiator Behavior Under the Threat of Arbitration
Discussants: ROBERT GIBBONS, Massachusetts Institute of Technology
CHARLES PLOTT, California Institute of Technology

10:15 A.M. TIME-SERIES ANALYSES OF MACROECONOMIC FLUCTUATIONS
Presiding: BEN S. BERNANKE, Princeton University
Papers: JOHN Y. CAMPBELL, Princeton University
Permanent and Transitory Components in Macroeconomic Fluctuations
MATTHEW D. SHAPIRO, Yale University
Are Fluctuations in Output Due More to Supply Shocks or Demand Shocks?

Discussants: JAMES H. STOCK, Harvard University
KENNETH D. WEST, Princeton University
STEPHEN P. ZELDES, University of Pennsylvania

10:15 A.M. ECONOMIC STATUS OF MINORITIES
Presiding: GEORGE J. BORJAS, University of California-Santa Barbara
Papers: CLAUDIA GOLDIN, University of Pennsylvania, AND SOLOMON W. POLACHEK, State University of New York-Binghamton
Perspectives on the Gender Gap
JAMES P. SMITH, Rand Corporation, AND FINIS R. WELCH, University of California-Los Angeles
Race and Poverty: A Forty-Year Record
Discussants: FRANCINE D. BLAU, University of Illinois
GLENN C. LOURY, Harvard University
SHERWIN ROSEN, University of Chicago

10:15 A.M. AFTER A HALF-CENTRY, WHAT VALID LESSONS REMAIN FROM KEYNES'S *GENERAL THEORY*?
Presiding: PAUL DAVIDSON, Rutgers University
Papers: PAUL DAVIDSON, Rutgers University
Sensible Expectations and the Long-Run Nonneutrality of Money
VICTORIA CHICK, University College, London
Speculation, the Rate of Interest, and Profit
JAN A. KREGEL, Johns Hopkins University Center, Bologna, Italy
The Effective Demand approach to Employment and Inflation Analysis
Discussants: WARREN J. SAMUELS, Michigan State University
JOHN CORNWALL, Dalhousie University, Canada
EILEEN R. APPELBAUM, Temple University

10:15 A.M. *LDC* TRADE AND FINANCIAL LIBERALIZATION IN RETROSPECT
Presiding: W. MAX CORDEN, Harvard University
Papers: MICHAEL MICHAELI, DEMETRIS PAPAGEORGIOUS, AND ARMEANE CHOKSI, World Bank
The Timing and Sequencing of Trade Liberalization Policy
SEBASTIAN EDWARDS, University of California-Los Angeles
On the Order of Liberalizing Current and Capital Accounts: Lessons from Latin America
RONALD I. MCKINNON, Stanford University
Interest Rates and Foreign Exchange Management in a Liberalizing Economy
Discussants: JEFFREY A. FRANKEL, University of California-Berkeley
ALBERT FISHLOW, University of California-Berkeley
MARCELO SELOWSKY, World Bank

10:15 A.M. CONCENTRATION AND PRICE, NOT CONCENTRATION AND PROFITS
Presiding: LEONARD W. WEISS, University of Wisconsin-Madison
Papers: CHRISTINA M. L. KELTON, Wayne State University
Change in Concentration, Change in Cost, and Change in Price—A Simultaneous Equation Approach
ROLAND KOLLER II, Brigham Young University
Concentration and Price in Regional Cement Markets
SCOTT MILLIMAN, JAMES MADISON UNIVERSITY, AND STEVEN T. BERRY, University of Wisconsin-Madison
The Effects of Concentration, Entry, and Entry Conditions on Unregulated Airline Fares
Discussants: RICHARD L. SCHMALENSEE, Massachusetts Institute of Technology
SAM PELTZMAN, University of Chicago

10:15 A.M. DYNAMIC GAMES
Presiding: HUGO SONNENSCHEIN, Princeton University
Papers: DILIP ABREU, Harvard University, AND ARIEL RUBINSTEIN, Hebrew University of Jerusalem
The Structure of Nash Equilibria in Repeated Games with Finite Automata
SANFORD GROSSMAN, Princeton University
Further Results on Sequential Bargaining under Asymmetric Information
DREW FUDENBERG, University of California-Berkeley, AND PAUL MILGROM, Yale University
Repeated Moral Hazard
FARUK GUL, Stanford University
Bargaining Foundations of Shapley Value

NOTES

The annual meeting of the American Economic Association will be held in New Orleans, Louisiana, December 28–30, 1986.

The Professional Placement Service will be located at the Sheraton New Orleans Hotel. It will be open from 10:00 A.M. to 5:00 P.M., December 27; 9:00 A.M. to 5:00 P.M., December 28–29; and 9:00 A.M. to 12:00 noon, December 30.

Members wishing to give papers or make suggestions for the program for the AEA meeting, Chicago, IL, December 28–30, 1987, are invited to write to Professor Robert Eisner, Department of Economics, Northwestern University, Evanston, IL 60201. To be considered for contributed sessions, abstracts of (noneconometric) papers must be received no later than February 1, 1987.

The Institute of International Education has awarded a grant to the American Economic Association under the Short-Term Enrichment Program (STEP) of the U.S. Information Agency to assist foreign graduate students studying in the United States to attend the annual meeting of the Association. Recipients must be full-time graduate students at U.S. institutions of higher learning and must not be receiving *any* U.S. government funds for academic or travel expenses. The 1986 meetings will be held in New Orleans, LA, December 28–30. The maximum award is $300. To receive an application form, write STEP Grants, American Economic Association, 1313 21st Avenue South, Suite 809, Nashville, TN 37212–2786. To be considered for a travel grant, completed forms must be received by November 3, 1986.

The Woodrow Wilson Center for Scholars seeks project proposals in the humanities and social sciences. Eligibility for academic participants is limited to the postdoctoral level; for participants from other backgrounds, equivalent professional achievement levels are expected. Fellows will devote full time on one major project while in residence at the Center for periods of 4 months to a year—stipends are subject to a ceiling of $35,000 per 12-month period. The deadline for receipt of applications is October 1 for appointments to begin the following September. For information and application materials, contact The Wilson Center, Smithsonian Institution Bldg, Rm 331, Washington, D.C. 20560 (telephone 202 + 357–2841).

The Columbia Society of Fellows in the Humanities, with grants from the Andrew W. Mellon Foundation and the William R. Kenan Trust, will appoint a number of postdoctoral fellows in the humanities for the academic year 1987–88. The appointment carries with it the expectation of renewal for a second year. Fellows newly appointed for 1987–88 must have received the Ph.D. between January 1, 1985 and July 1, 1987. The stipend will be $26,000, one-half for independent research and one-half for teaching in the undergraduate program in general education. Additional funds are available to support research. Application forms can be obtained by writing to the Director, Society of Fellows in the Humanities, Heyman Center for the Humanities, Box 100 Central Mail Room, Columbia University, New York, NY 10027. The deadline for receipt of completed application forms is October 15, 1986.

The Committee on Scholarly Communication (CSC-PRC) with the People's Republic of China announces a new five-year program. Proposals are invited for sustained research projects involving periodic visits of varying duration for 1987 to 1992 in Fengjia Village and Zouping County, Shandong Province. The CSCPRC will consider proposals such as 1) a single research proposing a discrete topic; 2) a research director either working with, or willing to incorporate, other individuals projects as appropriate into the field site; 3) a multidisciplinary team of three to five scholars who wish to pursue an integrated research project. In considering all proposals, the CSCPRC will give preference to projects that include prior field research experience in China, or in other countries with similar research conditions, and awareness of the Chinese research environment. For group projects, preference will be given to multidisciplinary proposals that include training of younger scholars, competence in aspects of rural development in other countries, and the commitment of at least one person to learn the local dialect. For further information, write the CSCPRC, National Academy of Sciences, 2101 Constitution Avenue, Washington, D.C. 20418, or telephone 202 + 334–2718. The deadline for applications is October 1, 1986.

The Division of International Programs of the National Science Foundation announces their U.S.–China Cooperative Program to promote opportunities for cooperation between scientists of the United States and the People's Republic of China on projects of mutual interest and benefit. The program supports joint research projects of up to two years' duration in all fields normally supported by NSF. Proposals should follow format and budgetary guidelines provided in the announcement (NSF 82–50: U.S.–China Cooperative Sci-

ence Program) which may be obtained from the Division of International Programs, NSF, 1800 G Street, N.W., Rm 1214, Washington, D.C. 20550 (telephone 202+357-7393).

The National Convention of the American Association for the Advancement of Slavic Studies, hosted by the Southern Conference on Slavic Studies, will be held November 20-23, 1986, at the Hyatt Regency Hotel, New Orleans, LA. For further information, contact AAASS, 128 Encina Commons, Stanford University, Stanford, CA 94304 (telephone 415+723-9668).

The American Statistical Association, under a grant from the National Science Foundation, announces a research fellowship and associate program in cooperation with the Bureau of Labor Statistics. Selected fellows and trainees will work at the BLS for periods of up to one year on related BLS programs that include general areas such as conceptual issues, measurement of nonwage benefits, measurement of economic growth, productivity-related research, labor market analyses, as well as survey methodology, statistical quality control, and development of time-series methods. Application deadline is January 15, 1987 for fellows; February 15, 1987 for associates. To obtain full information, write to Dr. Fred C. Leone, Executive Director, American Statistical Association, 806 15th Street, N.W., Washington, D.C. 20005.

The National Institutes of Health (NIH) Working Group on Health and Behavior announces the NIH National Research Service Award (NHRA) program. Fellowships are available to postdoctoral and senior scientists for up to three years. The NIH fellowships include research on basic processes as well as research applied to behavioral factors in health and illness. Receipt dates for applications are January 10, May 10, and September 10 of each year. For full information and applications, contact Office of Grant Inquiries, Division of Research Grants, NIH, Westwood Bldg 449, Bethesda, MD 20892 (telephone 301+496-7441).

The National Humanities Center will admit approximately forty fellows for the academic year, 1987-88. Fellowships are awarded on the basis of open competition, and applications are welcomed from the United States and abroad. Representatives of the humanities, natural sciences, and professional life may apply. The deadline is October 15, 1986. Application forms are available from the Center and should be supported by a curriculum vitae, a 1,000-word project proposal, and three letters of recommendation. Please write Kent Mullikin, Assistant Director, National Humanities Center, 7 Alexander Drive, Research Triangle Park, NC 27709.

Harvard Law School offers four or five fellowships for 1987-88 to college and university teachers in the social sciences and humanities to enable them to study fundamental techniques, concepts, and aims of law so that, in their teaching and research, they will be better able to use legal materials and legal insights which are relevant to their own disciplines. The deadline is January 15, 1987. Further information may be obtained from the Chairman, Committee on Liberal Arts Fellowships in Law, Harvard Law School, Cambridge, MA 02138.

The Leonard J. Savage Award of $500 is presented annually for an outstanding doctoral dissertation in Bayesian Econometrics and Statistics. To be considered for an award, the dissertation supervisor should submit the dissertation and a letter summarizing the main results. Dissertations completed after January 1, 1977, are eligible. The closing date each year is September 1. Send submissions to Professor Arnold Zellner, Graduate School of Business, 1101 East 58th Street, Chicago, IL 60637.

The winner of the 1985 award is Peter Jamison Lenk, "Bayesian Nonparametric Predictive Distributions," completed at the Department of Economics, University of Michigan.

The Harry J. Benda Prize of the Association of Asian Studies is awarded biennially to an outstanding young scholar in any field or country specialization of Southeast Asian studies. There is neither a citizenship nor residence requirement for nominees. Nominations for the 1987 prize should be sent before November 15, 1986 and include: name of nominee, affiliation, field of specialization, list of significant works, a brief statement about why she or he merits the award. Address: John A. Larkin, History Department, SUNY/Buffalo, Buffalo, NY 14261.

The Economics Department and the School of Natural Resources of the University of Michigan offer a joint Ph.D. Program in Natural Resource Economics. In the Economics Department, students pursue a rigorous program in economic theory, quantitative methods, resource economics, and another chosen field in economic analysis. In the School of Natural Resources (or elsewhere in the university), they are trained in the technical, biophysical aspects of a resource or environmental problem area of their choosing. For further information, please write to Professor Richard C. Porter, Director, Natural Resource Economics Program, Economics Department, Lorch Hall, University of Michigan, Ann Arbor, MI 48109.

The Social Science Research Council (SSRC) announces 1986 deadlines for 1987 academic year research fellowships and grants. *Predoctoral and Doctoral Dissertation Programs*: *November 3*: International Doctoral Research in Africa, Korea, Latin America and the Caribbean, Near and Middle East, South Asia, Southeast Asia, and Western Europe; *December 1*: Graduate Training in Soviet Studies—3rd and 4th year graduate study; Dissertation Fellowships—final year's work on dissertation. *Institutional Support Programs*: *December 1*: Program to Initiate New Teaching Positions in Russian and Soviet Studies. *Advanced Research Programs*: *December 1*: Grants for International Research in Africa, Japan, Korea, Latin America and the Caribbean, Near and Middle East, South Asia, and Southeast Asia; Advanced Research Fellowships in Foreign Policy Studies—1 or 2 year's support; Grants for Indochina Studies on Cambodia, Laos, and Vietnam; Advanced Research Grants for the Comparative Study of Muslim Societies; Advanced Grants in Soviet Studies for 3 summers and 1 semester of research. For details and how to apply, address the specific program at the SSRC. 605 Third Avenue, New York, NY 10158.

The deadline *December 1* is for the ACLS Fellowships (or Grants) for International Research in China and Eastern Europe: support in the social sciences and humanities. For details and how to apply, address the specific program at the American Council of Learned Societies, 228 East 45th Street, New York, NY 10017.

Call for Papers: The Bank Administration Institute and the Banking Research Center of the Kellogg Graduate School of Management, Northwestern University, will sponsor a conference on Asset Securitization and Off-Balance Sheet Risks of Depository Financial Institutions to be held at the Allen Center, Evanston Campus, Northwestern University, February 15–17, 1987. The proceedings will be published in *Studies in Banking and Finance*. Submit completed papers or abstracts to Professor Stuart I. Greenbaum, Banking Research Center, Kellogg GSM, Northwestern University, Evanston, IL 60201 (telephone 312+491–5498).

Call for Papers: The International Atlantic Economic Conference, "Creating New Traditions," will be held in Munich, West Germany, April 20–27, 1987. To present a paper, send two copies of a 500-word summary and cover page giving conference title, name, affiliation, mailing address and telephone number, appropriate *JEL* category, and $35 (US) submission fee. Send to John M. Virgo, International Program Chairman, AEC, Southern Illinois University, Box 1101, Edwardsville, IL 62026–1101.

To be held in conjunction is the International Health Economics and Management Conference. The same dates and guidelines apply for paper presentation ($50 fee). Accepted authors to provide personal expenses and registration fee. Send to John M. Virgo (above address: noting International Health...).

Call for Papers: The Hong Kong Economic Association will sponsor an international conference, "Economic Development in Chinese Societies: Models and Experiences," to be held in Hong Kong, December 18–21, 1986. The working languages are English and Chinese. To contribute or to attend, please contact Dr. Y. C. Jao, Department of Economics, University of Hong Kong, Pokfulam Road, Hong Kong.

Call for Papers: The 1987 annual meeting of the Business History Conference will be held March 12–14, in Wilmington, Delaware. The theme is "The History of International Business." Those interested in presenting a paper or developing a session should send a one-page précis and curriculum vitae to Professor Mira Wilkins, Department of Economics, Florida International University, Miami, FL 33199, before October 15, 1986.

Call for Papers: The annual meeting of the Eastern Finance Association will be held April 22–25, 1987, in Baltimore, Maryland. To present a paper, submit a two-page, single spaced abstract no later than November 1, 1986, or to chair a session, organize a special session, or serve as a discussant, write to Gershon N. Mandelker, Vice-President, Program, EFA, Graduate School of Business, Mervis Hall, University of Pittsburgh, PA 15260.

Call for Papers: The annual meeting of the Southern Regional Science Association will be held in Atlanta, Georgia, March 26–28, 1987. To present a paper, submit a one-page abstract no later than November 1, 1986, or to chair a session, organize symposia, or be a discussant, contact Professor William Latham, Department of Economics, University of Delaware, Newark, DE 19711 (telephone 302+451–2564).

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to airfare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its

equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for application to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 228 East 45 Street, New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

Retirements

Harvey E. Brazer: professor of economics, University of Michigan, December 31, 1985.

John E. Perkins, Jr.: professor of economics, Texas Christian University, May 1986.

W. M. Scammell: professor of economics, McMaster University.

Foreign Scholars

Luis Rene Caceres, Central American Bank of Economic Integracion, Tegucigalpa, Honduras: visiting professor, department of economics, Florida International University, January 1986.

Mechthild Minkner, Institut Fur Iberoamerika Kunde, Hamburg, Germany: visiting research professor, department of economics, Florida State University, January 1986.

Dodli Sai Prasado-Rao, University of New England, Armidale, Australia: visiting professor, department of economics, Florida International University, January 1986.

Promotions

G. J. Anderson: professor, department of economics, McMaster University, July 1986.

Harold S. Beebout: senior vice president and director of research, Mathematica Policy Research, Inc., February 28, 1986.

Steven D. Braithwait: senior project manager, Demand-Side Planning Program, Electric Power Research Institute, November 1, 1986.

M. J. Browning: professor, department of economics, McMaster University, July 1986.

Andrew Buck: associate professor, department of economics, Temple University, September 1985.

Ahmad Faruqui: manager, End-Use Assessment and Forecasting Subprogram, Electric Power Research Institute, November 1, 1985.

Michael Goetz: associate professor, department of economics, Temple University, September 1985.

Roger Gordon: professor, department of economics, University of Michigan, September 1, 1986.

Philip Gregorowicz: associate professor of economics, Auburn University-Montgomery, September 1986.

Sydney S. Hicks: senior vice president and chief economist, InterFirst Bank Dallas, July 1, 1985.

David Marwick: group director/transportation, U.S. General Accounting Office, October 1985.

S. Mestelman: professor, department of economics, McMaster University, July 1986.

Charles E. Metcalf: president, Mathematica Policy Research, Inc., February 28, 1986.

Arnoldt J. Spitz: executive vice president, finance and administration, International Seaway Trading Corporation, February 13, 1986.

Administrative Appointments

Joseph Horton: dean, School of Management, University of Scranton, June 1, 1986.

Michael S. McPherson: chair, economics department, Williams College, July 1, 1986.

David R. Meinster: chairman, economics department, Temple University, March 1986.

Richard C. Porter: chair, department of economics, University of Michigan, January 1, 1986.

Frederic E. Wakeman, Jr.: president, Social Science Research Council, New York, July 1, 1986.

Tom S. Witt: executive director, Bureau of Business Research and director, Center for Economic Analysis and Statistics, West Virginia University, January 1, 1986.

William C. Wood: chairman, department of economics and business administration, Bridgewater College, September 1, 1986.

Appointments

Harold Black: visiting professor, American University, September 1985.

Robert A. Blecker: instructor, American University, September 1985.

John Bound, Harvard University: assistant professor, department of economics, University of Michigan, September 1, 1986.

E. Kwan Choi, University of Missouri: associate professor, department of economics, Iowa State University, June 1, 1985.

Jens Christiansen, University of Massachusetts-Amherst: lecturer in economics, Mount Holyoke College, fall 1986.

M. Keivan Deravi: assistant professor of economics, Auburn University of Montgomery, September 1985.

Mark E. Edelman, South Dakota State University: associate professor, department of economics, Iowa State University, January 1, 1986.

Yiu-Kwan Fan, University of Wisconsin-Stevens Point: dean, faculty of business, Hong Kong Baptist College, September 1, 1986.

Paul Flacco: assistant professor, American University, September 1985.

Charles E. Hegji: assistant professor of economics, Auburn University-Montgomery, June 1985.

Michael W. Horrigan, Williams College: research economist, Division of Employment and Unemployment Analysis, Bureau of Labor Statistics, 1986.

Richard P. Hydell, Wake Forest University: assistant professor of economics, Mary Washington College, August 1986.

Helen Jensen, University of Maryland: assistant professor, department of economics, Iowa State University, August 21, 1985.

Lori Gladstein Kletzer, University of California-Berkeley: assistant professor of economics, Williams College, July 1, 1986.

Jeffrey K. Mackie-Mason, MIT: assistant professor, department of economics, University of Michigan, September 1, 1986.

Jon J. Manger, La Salle University: international economist, U.S. Department of Commerce, February 1986.

James C. Murdock: assistant professor of economics, Auburn University-Montgomery, June 1986.

Eva Paus, University of Pittsburgh: assistant professor of economics, Mount Holyoke College, January 1987.

Irene Powell: assistant professor of economics, Mount Holyoke College, fall 1986.

John Rizzo, American Medical Association: adjunct assistant professor of economics and research scientist, Center of Japan–U.S. Business and Economic Studies, New York University, February 1986.

David R. Ross, Williams College: visiting assistant professor of economics, University of Pennsylvania, July 1986–June 1987.

Efraim Sadka, Tel Aviv University: visiting professor, department of economics, University of Michigan, September 1, 1986.

Ryuzo Sato, Brown University: professor of economics and director, Center for Japan–U.S. Business and Economic Studies, New York University, September 1985.

Allen L. Schirm, University of Pennsylvania: visiting assistant professor, department of economics, University of Michigan, September 1, 1986.

Joseph Swierzbinski, University of Washington: visiting assistant professor, department of economics, University of Michigan, September 1, 1986.

Jerry Thursby, Ohio State University: visiting associate professor, department of economics, University of Michigan, September 1, 1986.

M. R. Veall: assistant professor, department of economics, McMaster University, July 1986.

Robert Vogel, World Bank: visiting professor, American University, September 1986.

Marina von Neumann Whitman, General Motors Company: adjunct professor of economics, business, and public policy, department of economics, University of Michigan, September 1, 1986.

Christine Wong, Mount Holyoke College: assistant professor of economics, University of California-Santa Cruz.

Huizhong Zhou; Northwestern University: visiting assistant professor, department of economics, University of Michigan, September 1, 1986.

Leave for Special Appointments

Lee J. Alston, Williams College: visiting fellow, department of economic history, Research School of Social Sciences, Australian National University, July 1986–July 1987.

Marsha E. Courchane, North Carolina State University: visiting assistant professor, University of British Columbia, Vancouver, BC, July 1, 1986–June 30, 1987.

David Fairris, Williams College: visiting scholar, department of economics and Institute of Industrial Relations, University of California-Berkeley, July 1986–June 1987.

Robert M. Fearn, North Carolina University: visiting professor, Higher School of Economics and Business, Athens, Greece, August 15, 1986–August 14, 1987.

Roger Gordon, University of Michigan: People's University, Beijing, China, September 1, 1986–December 31, 1986.

Edward Gramlich, University of Michigan: deputy director, Congressional Budget Office, January 1, 1986–July 1, 1987.

Brian Levy, Williams College: visiting research associate, Harvard Institution for International Development, July 1986–June 1987.

Maria Thursby, Ohio State University: NSF Program of Visiting Professorships for Women, University of Michigan, 1986–87.

Michelle White, University of Michigan: People's University, Beijing, China, September 1, 1986–December 31, 1986.

AMERICAN ECONOMIC ASSOCIATION

1986 ANNUAL MEMBERSHIP RATES

Membership includes:

—a subscription to both *The American Economic Review* (quarterly) plus *Papers and Proceedings* and the *Journal of Economic Literature* (quarterly).

- Regular members with annual incomes of $30,000 or less $37.50
- Regular members with annual incomes above $30,000 but no more than $40,000 .... $45.00
- Regular members with annual incomes above $40,000 .... $52.50
- Junior members (available to registered students for three years only). Student status must be certified by your major professor or school registrar $18.75
- In Countries other than the U.S.A., Add $12.00 to cover postage.
- Family members (persons living at the same address as a regular member, additional memberships without subscription to the publications of the Association) $7.50

Please begin my issues with:

☐ **March** ☐ **June** (Includes Papers and Proceedings) ☐ **September** ☐ **December**

First Name and Initial ____ Last Name ____ Suffix ____

Address Line 1 ____

Address Line 2 ____

City ____

State or Country ____ Zip/Postal Code ____

MAJOR FIELDS (TWO ONLY)
LIST FIELDS WITH WHICH YOU CURRENTLY IDENTIFY. SELECT FIELD CODE FROM *JEL*, "Classification System for Books."

☐ ☐

Please type or print information above; Please pay us either with a check or money order payable in United States Dollars. Canadian and foreign payments must be in the form of a draft or check drawn on a United States bank payable in United States Dollars. Please note: It is the policy of the Association, not to refund membership payments.

Endorsed by (AEA member) ____

Below for Junior Members Only

I certify that the person named above is enrolled as a student at ____

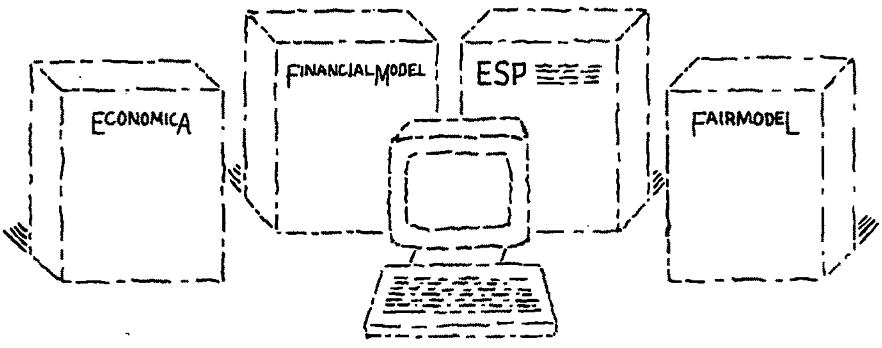
____

Authorized Signature ____

PLEASE SEND WITH PAYMENT TO:

AMERICAN ECONOMIC ASSOCIATION
1313 21ST AVENUE SOUTH, SUITE 809
NASHVILLE, TENNESSEE 37212-2786
U.S.A.

The First Family of Forecasting



ECONOMICA, Inc., the pioneer of PC-based forecasting, specializes in software for economists, managers, and planners. We've combined the speed and convenience of personal computing with the accuracy of advanced economic modeling to put the future at your fingertips.

No Delays, No Hidden Fees

Because ECONOMICA software operates on your own IBM® PC, PC/XT,™ PC AT,™ or any other MS-DOS™ system, we've eliminated the delays, constraints and hidden costs of mainframe services. You can generate forecasts as often as you like, for as many variables as you need. And you can customize the equations and alter the assumptions with just a few keystrokes.

Fully Integrated Software

ECONOMICA offers a fully integrated software family. This allows you to move data from one program to another without redundant keystroking or tedious programming. And all ECONOMICA programs present a variety of format options—from simple spreadsheets to elaborate graphics.

From the Makers of FAIRMODEL

To satisfy all your forecasting needs, ECONOMICA announces the addition of two software programs to its product line, both compatible with our acclaimed FAIRMODEL package for macroeconomic forecasting.

ESP®-The Econometric Software Package™

ECONOMICA now has publishing rights to ESP, the premier econometric software package. This comprehensive forecasting tool integrates econometric and statistical analyses with advanced graphics and data management. With ESP, you can formulate your own models and generate customized forecasts using either your own data or FAIRMODEL predictions—or both, so that you can monitor the effects of economic changes in your own industry and in your markets on your company's sales and profits.

FINANCIAL MODEL

Designed by Dr. Ray Fair, named the nation's top forecasting economist in the *Business Week* survey, Financial Model evaluates portfolio outlook through the year 2000. The model forecasts money supply, interest rates, flow of funds, deposits, outstanding credit, bond yields, Eurodollar rates, the Standard and Poor's 500-stock Index and more.

For more information, Call ECONOMICA, Inc., 2067 Massachusetts Ave., Cambridge, Mass. (617) 661-3260; TELEX via WUI 6502773397 MCI.

FROM THE GROWING ECONOMICA inc SOFTWARE FAMILY